Coevolution of Deception and Preferences: Darwin and Nash Meet Machiavelli

Heller, Yuval and Mohlin, Erik

Bar Ilan University, Lund University

12 February 2017

Coevolution of Deception and Preferences: Darwin and Nash Meet Machiavelli*

Yuval Heller† and Erik Mohlin‡

21st September 2018

Final pre-print of a manuscript accepted for publication in *Games and Economic Behavior*.

Abstract

We develop a framework in which individuals' preferences coevolve with their abilities to deceive others about their preferences and intentions. Specifically, individuals are characterised by (i) a level of cognitive sophistication and (ii) a subjective utility function. Increased cognition is costly, but higher-level individuals have the advantage of being able to deceive lower-level opponents about their preferences and intentions in some of the matches. In the remaining matches, the individuals observe each other's preferences. Our main result shows that, essentially, only efficient outcomes can be stable. Moreover, under additional mild assumptions, we show that an efficient outcome is stable if and only if the gain from unilateral deviation is smaller than the effective cost of deception in the environment.

Keywords: Evolution of Preferences; Indirect Evolutionary Approach; Theory of Mind; Depth of Reasoning; Deception; Efficiency. **JEL codes:** C72, C73, D03, D83.

Preprint of the

*Valuable comments were provided by the anonymous associate editor and referees, Vince Crawford, Eddie Dekel, Jeffrey Ely, Itzhak Gilboa, Christoph Kuzmics, Larry Samuelson, Jörgen Weibull, and Okan Yilankaya, as well as participants at presentations at Oxford University, Queen Mary University, G.I.R.L.13 in Lund, the Toulouse Economics and Biology Workshop, DGL13 in Stockholm, the 25th International Conference on Game Theory at Stony Brook, and the Biological Basis of Preference and Strategic Behaviour 2015 conference at Simon Fraser University. Yuval Heller is grateful to the *European Research Council* for its financial support (starting grant #677057). Erik Mohlin is grateful to *Handelsbankens forskningsstiftelser* (grant #P2016-0079:1) and the *Swedish Research Council* (grant #2015-01751) for its financial support.

†Affiliation: Department of Economics, Bar Ilan University. Address: Ramat Gan 5290002, Israel. E-mail: yuval.heller@biu.ac.il.

‡Affiliation: Department of Economics, Lund University. Address: Tycho Brahes väg 1, 220 07 Lund, Sweden. E-mail: erik.mohlin@nek.lu.se.

1 Introduction

For a long time economists took preferences as given. The study of their origin and formation was considered a question outside the scope of economics. Over the past two decades this has changed dramatically. In particular, there is now a large literature on the evolutionary foundations of preferences (for an overview, see Robson and Samuelson, 2011). A prominent strand of this literature is the so-called "indirect evolutionary approach," pioneered by Güth and Yaari (1992) (term coined by Güth, 1995). This approach has been used to explain the existence of a variety of "non-standard" preferences that do not coincide with material payoffs, e.g., altruism, spite, and reciprocal preferences.[1] Typically, the non-materialistic preferences in question convey some form of commitment advantage that induces opponents to behave in a way that benefits individuals with non-materialistic preferences, as described by Schelling (1960) and Frank (1987). Indeed, Heifetz, Shannon, and Spiegel (2007) show that this kind of result is generic.

A crucial feature of the indirect evolutionary approach is that preferences are explicitly or implicitly assumed to be at least partially observable.[2] Consequently the results are vulnerable to the existence of mimics who signal that they have, say, a preference for cooperation, but actually defect on cooperators, thereby earning the benefits of having the non-standard preference without having to pay the cost (Samuelson, 2001). The effect of varying the degree to which preferences can be observed has been investigated by Ok and Vega-Redondo (2001), Ely and Yilankaya (2001), Dekel, Ely, and Yilankaya (2007), and Herold and Kuzmics (2009). They confirm that the degree to which preferences are observed decisively influences the outcome of preference evolution.

Yet, the degree to which preferences are observed is still exogenous in these models. In reality we would expect both the preferences and the ability to observe or conceal them to be the product of an evolutionary process.[3] *This paper provides a first step towards filling in the missing*

[1] For example, Bester and Güth (1998), Bolle (2000), and Possajennikov (2000) study combinations of altruism, spite, and selfishness. Ellingsen (1997) finds that preferences that induce aggressive bargaining can survive in a Nash demand game. Fershtman and Weiss (1998) study evolution of concerns for social status. Sethi and Somanthan (2001) study the evolution of reciprocity in the form of preferences that are conditional on the opponent's preference type. In the context of the finitely repeated Prisoner's Dilemma, Guttman (2003) explores the stability of conditional cooperation. Dufwenberg and Güth (1999) study firm's preferences for large sales. Güth and Napel (2006) study preference evolution when players use the same preferences in both ultimatum and dictator games. Koçkesen, Ok, and Sethi (2000) investigate survival of more general interdependent preferences in aggregative games. Friedman and Singh (2009) show that vengefulness may survive if observation has some degree of informativeness. Recently, Norman (2012) has shown how to adapt some of these results into a dynamic model

[2] Gamba (2013) is an interesting exception. She assumes play of a self-confirming equilibrium, rather than a Nash equilibrium, in an extensive-form game. This allows for evolution of non-materialistic preferences even when they are completely unobservable. An alternative is to allow for a dynamic that is not strictly payoff monotonic. This approach is pursued by Frenkel, Heller, and Teper (forthcoming), who show that multiple biases (inducing non-materialistic preferences) can survive in non-monotonic evolutionary dynamics even if they are unobservable, because each approximately compensates for the errors of the others.

[3] On this topic, Robson and Samuelson (2011) write: "The standard argument is that we can observe preferences because people give signals – a tightening of the lips or flash of the eyes – that provide clues as to their feelings. However, the emission of such signals and their correlation with the attendant emotions are themselves the product of evolution. [...] We cannot simply assume that mimicry is impossible, as we have ample evidence of mimicry from

link between evolution of preferences and evolution of how preferences are concealed, feigned, and detected.[4] In our model the ability to observe preferences and the ability to deceive and induce false beliefs about preferences are endogenously determined by evolution, jointly with the evolution of preferences. Cognitively more sophisticated players have positive probability of deceiving cognitively less sophisticated players. Mutual observation of preferences occurs only in matches in which such deception fails. This setup is general enough to encompass both the standard indirect evolutionary model where preferences are always observed, and the reverse case in which more sophisticated types always deceive lower types, as well as all intermediate cases between these two extremes. *We find that, generically, only efficient outcomes can be played in stable population states.* Moreover, we define a single number that captures the effective cost of deception against naive opponents, and show that *an efficient outcome is stable if and only if the gain from a unilateral deviation is smaller than the effective cost of deception.*

Overview of the Model. As is common in standard evolutionary game theory we assume an infinite population of individuals who are uniformly randomly matched to play a symmetric normal form game.[5] Each individual has a type, which is a tuple, consisting of a *preference component* and a *cognitive component*. The preference component is identified with a subjective utility function over the set of outcomes (i.e. action profiles), which may differ from the objective payoffs (i.e., fitness) of the underlying game. The cognitive component is simply a natural number representing the level of cognitive sophistication of the individual.[6,7] The cost of increased cognition is strictly positive.

When two individuals with different cognitive levels are matched, there is positive probability (which may depend on the cognitive levels of both agents) that the agent with the higher level deceives his opponent. For the sake of tractability, and in order not to limit the degree to which

the animal world, as well as experience with humans who make their way by misleading others as to their feelings, intentions and preferences. [...] In our view, *the indirect evolutionary approach will remain incomplete until the evolution of preferences, the evolution of signals about preferences, and the evolution of reactions to these signals, are all analysed within the model.*" [Emphasis added] (pp. 14–15)

[4]The recent working paper of Gauer and Kuzmics (2016) presents a different way to endogenising the observability of preferences. Specifically, they assume that preferences are ex ante uncertain, and that each player may exert a cognitive effort to privately observe the opponent's preferences.

[5]It is known that positive assortative matching is conducive to the evolution of altruistic behaviour (Hines and Maynard Smith, 1979) and non-materialistic preferences even when preferences are perfectly unobservable (Alger and Weibull, 2013; Bergstrom, 1995). It is also known that finite populations allow for evolution of spiteful behaviours (Schaffer, 1988) and non-materialistic preferences (Huck and Oechssler, 1999). By assuming that individuals are uniformly randomly matched in an infinite population, we avoid confounding these effects with the effect of endogenising the degree of observability.

[6]The one-dimensional representation of cognitive ability reflects the idea that if one is good at deceiving others, then one is more likely to be good also at reading others and avoiding being deceived by them. In this paper we simplify this relation by assuming a perfect correlation between the two abilities, and leave the study of more general relations for future research.

[7]Remark 7 in Section 2.2 presents an alternative interpretation of our model, according to which this cognitive component represents the agent's social status, rather than the agent's ability to deceive other agents.

higher levels may exploit lower levels, we model a strong form of deception. The deceiver observes the opponent's preferences perfectly, and is allowed to choose whatever she wants the deceived party to believe about the deceiver's intended action choice. A strategy profile that is consistent with this form of deception is called a *deception equilibrium*. With the remaining probability (or with probability one if both agents have the same cognitive level) there is no deception in the match. In this case, we assume that each player observes the opponent's preferences, and the individuals play a Nash equilibrium of the complete information game induced by their subjective preferences.

The state of a population is described by a *configuration*, consisting of a type distribution and a behaviour policy. The *type distribution* is simply a finite support distribution on the set of types. The *behaviour policy* specifies a Nash equilibrium for each match without deception, and a deception equilibrium for each match with deception. In a *neutrally stable configuration* all incumbents earn the same expected fitness, and if a small group of mutants enter they earn weakly less than the incumbents in any *focal* post-entry state. A focal post-entry state is one in which the incumbents behave against each other in the same way as before the mutants entered.

Main Results. We say that a strategy profile is (fitness-)efficient if it maximises the sum of objective payoffs. Theorem 1 shows that in any stable configuration, any type $\bar{\theta}$ with the highest cognitive level in the incumbent population must play an efficient strategy profile when meeting itself. The intuition is that otherwise a highest-type mutant who mimics the play of $\bar{\theta}$ against all incumbents while playing an efficient strategy profile against itself would outperform type $\bar{\theta}$ (a novel application of the "secret handshake" argument due to Robson, 1990).

Next we restrict attention to generic games (i.e. games that result with probability one if fitness payoffs are independently drawn from a continuous distribution) and obtain our first main result: *any stable configuration must induce efficient play* in all matches between all types. The idea of the proof can be briefly sketched as follows. We first show that any type θ in a stable configuration must play an efficient strategy profile when meeting *itself*. Otherwise a mutant who has the same level as θ and the same utility function as θ, but who plays efficiently against itself, could invade the population. Next, we show that *any* two types must play an efficient strategy profile. The intuition is that otherwise the average within-group fitness would be higher than the between-group fitness, which implies instability in the face of small perturbations in the frequency of the types: a type who became slightly more frequent would have a higher fitness than the other incumbents, and this would move the population away from the original configuration.

The existing literature (e.g., Dekel, Ely, and Yilankaya, 2007) has demonstrated that if players perfectly observe each other's preferences (or do so with sufficiently high probability), then only efficient outcomes are stable. As was pointed out above, our model encompasses the limiting case in which it is arbitrarily "cheap and easy" to deceive the opponent, i.e. the case in which the

marginal cost of an additional cognitive level is very low, and having a slightly higher cognitive level allows a player to deceive the opponent with probability one. A key contribution of the paper is to show that even when it is cheap and easy to deceive the opponent, *the seemingly mild assumption of perfect observability, and Nash equilibrium behaviour, among players with the same cognitive level is enough to ensure that stability implies efficiency.*

In order to obtain sufficient conditions for stability we restrict attention to generic games that admit a "punishment action" that ensures that the opponent achieves strictly less than the symmetric efficient fitness payoff. For games satisfying this relatively mild requirement we fully characterise stable configurations. We define the *(fitness) deviation gain* of an action profile to be the maximal fitness increase a player may obtain by unilaterally deviating from this action profile (this gain is zero if and only if the action profile is a Nash equilibrium of the underlying game). Next we define the *effective cost of deception* in the environment as the minimal ratio between the cost of an increased cognitive level and the probability that an agent with this level deceives an opponent with the lowest cognitive level. Our second main result shows that an efficient action profile is the outcome of a stable configuration if and only if its deviation gain is smaller than the effective cost of deception. In particular, *efficient Nash equilibria are stable in all environments, while non-Nash efficient action profiles are stable only as long as the gain from a unilateral deviation is sufficiently small.*

Next, we note that non-generic games may admit different kinds of stable configurations. One particularly interesting family of non-generic games is the family of zero-sum games, such as the Rock-Paper-Scissors game. We analyse this game and characterise a heterogeneous stable population (inspired by a related construction in Conlisk, 2001) in which different cognitive levels coexist, players with equal levels play the Nash equilibrium of the underlying game, and players with higher levels beat their opponents but this gain is offset by higher cognitive costs.

Finally, in Section 4 we discuss two extensions of the model (which are formally analysed in Appendices B and D): (1) we relax the assumption that each agent perfectly observes the partner's preferences in matches without deception, and (2) we allow for type-interdependent preferences (à la Herold and Kuzmics, 2009), which are represented by utility functions that are defined over both action profiles and the opponent's type.

Further Related Literature. Our model is related to work in biology and evolutionary psychology on the evolution of the "theory of mind" (Premack and Woodruff, 1979), specifically, the "Machiavellian intelligence" hypothesis (Humphrey, 1976) and the "social brain" hypothesis (Byrne and Whiten, 1998), according to which the extraordinary cognitive abilities of humans evolved as a result of the demands of social interactions, rather than the demands of the natural environment: in a single-person decision problem there is a fixed benefit from being smart, but in a strategic situation it may be important to be smarter than the opponent. From an evolutionary perspective,

there is a trade-off between the benefit of outsmarting the opponent and the non-negligible costs associated with increased cognitive capacity (Holloway, 1996; Kinderman, Dunbar, and Bentall, 1998). Our model incorporates these features.

There is a smaller literature on the evolution of strategic sophistication within game theory; see, e.g., Stahl (1993), Banerjee and Weibull (1995), Stennek (2000), Conlisk (2001), Abreu and Sethi (2003), Mohlin (2012), Rtischev (2016), and Heller (2015). Following these papers, we provide results to the effect that different degrees of cognitive sophistication may coexist.

Robalino and Robson (2016) construct a model to demonstrate the advantage of having a theory of mind (understood as an ability to ascribe stable preferences to other players) over learning by reinforcement. In novel games the ascribed preferences allow the agents with a theory of mind to draw on past experience whereas a reinforcement learner without such a model has to start over again. Hopkins (2014) explains why costly signaling of altruism may be especially valuable for those agents who have a theory of mind.

Robson (1990) initiated a literature on evolution in cheap-talk games by formulating the secret handshake effect: evolution selects an efficient stable state if mutants can send messages that the incumbents either do not see or do not benefit from seeing. Against the incumbents a mutant plays the same action as the incumbents do, but against other mutants the mutant plays an action that is a component of the efficient equilibrium. Thus the mutants are able to invade unless the incumbents are already playing efficiently. See also the related analysis in Matsui (1991) and Wiseman and Yilankaya (2001). We allow for deception and still find that efficiency is necessary (though no longer sufficient) for stability. As pointed out by Wärneryd (1991) and Schlag (1993), among others, problems arise if either the incumbents use all available messages (so that there is no message left for the incumbents to coordinate on) or the incumbents follow a strategy that induces the mutants to play an action that lowers the mutants' payoffs below those of the incumbents. To circumvent this problem, Kim and Sobel (1995) use stochastic stability arguments and Wärneryd (1998) uses complexity costs. Similarly, evolution selects an efficient outcome in our model, where the preferences also serve the function of messages.

We conclude by mentioning three other related strands of literature in which deception has been implicitly studied: (1) the "strategic teaching" literature, which studies situations in which sophisticated agents manipulate the learning input of opponents in order to change the beliefs and future actions of these opponents (see, e.g., Fudenberg and Levine, 1998; Camerer, Ho, and Chong, 2002; Schipper, 2017, Section 8.11); (2) the "reputation" literature, in which a long-run player manipulates the beliefs and behaviour of short-run opponents (see Mailath and Samuelson, 2006, for a textbook exposition); and (3) non-equilibrium level-k analysis of games of conflict, where agents can use pre-play communication to deceive naive opponents (see, e.g., Crawford, 2003).

Structure. The rest of the paper is organised as follows. Section 2 presents the model. The results are presented in Section 3. In Section 4 we extend the model to deal with partial observability (formally analysed in Appendix D) and type-interdependent preferences (formally analysed in Appendix B). We conclude in Section 5. Appendix A contains proofs not in the main text. Appendix C formally constructs heterogeneous stable populations in specific games.

2 Model

We consider a large population of agents, each of whom is endowed with a type that determines her subjective preferences and her cognitive level. The agents are randomly matched to play a symmetric two-player game. A dynamic evolutionary process of cultural learning, or biological inheritance, increases the frequency of more successful types. We present a static solution concept to capture stable population states in such environments.

2.1 Underlying Game and Types

Consider a symmetric two-player normal form game G with a finite set A of pure actions and a set $\Delta(A)$ of mixed actions (or strategies). We use the letter a (resp., σ) to describe a typical pure action (resp., mixed action). Payoffs are given by $\pi : A \times A \to \mathbb{R}$, where $\pi(a, a')$ is the material (or fitness) payoff to a player using action a against action a'. The payoff function is extended to mixed actions in the standard way, where $\pi(\sigma, \sigma')$ denotes the material payoff to a player using strategy σ, against an opponent using strategy σ'. With a slight abuse of notation let a denote the degenerate strategy that puts all the weight on action a. We adopt this convention for probability distributions throughout the paper.

Remark 1. Asymmetric interactions can be captured in our setup (as is standard in the literature; see, e.g., Brown and von Neumann, 1950; Selten, 1980; van Damme, 1987, Section 9.5) by embedding the asymmetric interaction in a larger, symmetric game in which nature first randomly assigns the players to roles in the asymmetric interaction.

We imagine a large population of individuals (technically, a continuum) who are uniformly randomly matched to play the game G. Each individual i in the population is endowed with a *type* $\theta = (u, n) \in \Theta = U \times \mathbb{N}$, consisting of *preferences*, identified with a von Neumann–Morgenstern utility function, $u \in U$, and *cognitive level*[8] $n \in \mathbb{N}$. Let $\Delta(\Theta)$ be the set of all finite support probability distributions on Θ. A population is represented by a finite-support *type distribution* $\mu \in \Delta(\Theta)$. [9] Let $C(\mu)$ denote the support (carrier) of type distribution $\mu \in \Delta(\Theta)$. Given a type

[8] For tractability, we choose to work with a discrete set of cognitive levels. The main results in the paper can be adapted to a setup in which the feasible set of cognitive efforts is a continuum, provided that we maintain our focus on finite-support type distributions.

[9] Comment 6 in Section 2.2 explains why we restrict attention to finite-support type distributions.

θ, we use u_θ and n_θ to refer to its preferences and cognitive level, respectively.

In the main model we assume that the preferences are defined over action profiles, as in Dekel, Ely, and Yilankaya (2007).[10] This means that any preferences can be represented by a utility function of the form $u : A \times A \to \mathbb{R}$. The set of all possible (modulo affine transformations) utility functions on $A \times A$ is $U = [0,1]^{|A|^2}$. Let $BR_u(\sigma')$ denote the set of best replies to strategy σ' given preferences u, i.e. $BR_u(\sigma') = \arg\max_{\sigma \in \Delta(A)} u(\sigma, \sigma')$.

There is a fitness cost to increased cognition, represented by a strictly increasing cognitive cost function $k : \mathbb{N} \to \mathbb{R}_+$ satisfying $\lim_{n\to\infty} k(n) = \infty$. The fitness payoff of an individual equals the material payoff from the game, minus the cognitive cost. Let k_n denote the cost of having cognitive level n. Hence $k_\theta = k_{n_\theta}$ denotes the cost of having type θ. Without loss of generality, we assume that $k_1 = 0$.

We would like to put forward two motivations for the assumption that there is an increasing fitness cost of having a higher cognitive level. The first motivation is relevant to settings in which the evolution of types is influenced by biological inheritance. There is a literature in biology and biological anthropology showing that brain volume, especially neocortex volume, is correlated with the size of social groups across species. Noting that brain tissue is metabolically costly, it has been argued that the size of the brain (in particular the neocortex) is at least partially determined by complexity of social organisation (see Dunbar, 1998, for a summary of the evidence and the arguments), which is in line with the "Machiavellian intelligence" and "social brain" hypotheses (Humphrey, 1976; Byrne and Whiten, 1997; Whiten and Byrne, 1988).

The second motivation is relevant also in setups in which types evolve as part of a social learning process. For concreteness, suppose that agents face two kinds of decision problems throughout their lives: (1) individual (ecological) decision problems against nature, and (2) interactive (social) decision problems as represented by playing the underlying game G. Agents have limited cognitive capacity. New agents who join the population face a trade-off between developing their deception-related cognitive skills (which are helpful when playing the game G) and developing other skills (which are helpful in the decision problems against nature). When a new agent joins the population, his type $\theta = (u_\theta, n_\theta)$ determines how much effort the agent exerts in developing his deception-related cognitive ability n_θ (while the remaining effort is exerted to develop the other skills). The increasing cognitive cost function $k(n_\theta)$ captures the agent's loss due to his sub-optimal performance in the decision problems against nature, which is induced by diverting effort to developing his deception-related cognitive ability at the expense of developing the other skills.

[10] In Appendix B, we study *type-interdependent* preferences, which depend on the opponent's type, as in Herold and Kuzmics (2009).

2.2 Configurations

A state of the population is described by a type distribution and a behaviour policy for each type in the support of the type distribution. An individual's behaviour is assumed to be (subjectively) rational in the sense that it maximises her subjective preferences given the belief she has about the opponent's expected behaviour. However, her beliefs may be incorrect if she is deceived by her opponent. An individual may be deceived if her opponent is of a strictly higher cognitive level. The probability of deception is given by the function $q : \mathbb{N} \times \mathbb{N} \to [0,1]$ that satisfies $q(n,n') = 0$ if and only if $n \leq n'$.[11] We interpret $q(n,n')$ as the probability that a player with cognitive level n deceives an opponent with cognitive level n'. Specifically, when two players with cognitive levels n' and $n \geq n'$ are matched to play, then with a probability of $q(n,n')$ the individual with the higher cognitive level n (henceforth, the *higher type*) observes the opponent's preferences perfectly, and is able to deceive the opponent (henceforth, the *lower type*). The deceiver is allowed to choose whatever she wants the deceived party to believe about the deceiver's intended action choice. The deceived party best-replies given her possibly incorrect belief. For simplicity, we assume that if the deceived party has multiple best replies, then the deceiver is allowed to break indifference, and choose which of the best replies she wants the deceived party to play. Consequently the deceiver is able to induce the deceived party to play any strategy that is a best reply to some belief about the opponent's mixed action, given the deceived party's preferences.

Given preferences $u \in U$, let $\Sigma(u)$ denote the set of *undominated strategies*, which are the set of actions that are best replies to at least one strategy of the opponent (given the preferences u). Formally, we define

$$\Sigma(u) = \{\sigma \in \Delta(A) : \text{ there exists } \sigma' \in \Delta(A) \text{ such that } \sigma \in BR_u(\sigma')\}.$$

We say that a strategy profile is a *deception equilibrium* if the strategy profile is optimal from the point of view of the deceiver under the constraint that the deceived player has to play an undominated strategy. Formally:

Definition 1. Given two types θ, θ' with $n_\theta > n_{\theta'}$, a strategy profile $(\tilde{\sigma}, \tilde{\sigma}')$ is a *deception equilibrium* if

$$(\tilde{\sigma}, \tilde{\sigma}') \in \arg \max_{\sigma \in \Delta(A), \sigma' \in \Sigma(u_{\theta'})} u_\theta(\sigma, \sigma').$$

Let $DE(\theta, \theta')$ be the set of all such deception equilibria.

With the remaining probability of $1 - q(n,n') - q(n',n)$ there is no deception between the

[11] One can extend our main results to a setup in which individuals with lower cognitive levels can deceive opponents with higher cognitive levels with a sufficiently small probability. Specifically, assume that for each generic game, there exists $\epsilon > 0$, such that $q(n,n') < \epsilon$ for each $n \leq n'$ (instead of requiring $q(n,n') = 0$). One can show that the characterization of NSCs in Corollary 2 remains qualitatively the same. Namely, the only candidates to be NSCs are configurations in which all agents have the minimal cognitive level, and all agents play the efficient action profile in every match with no deception. These configurations are NSCs if the effective cost of defection is sufficiently low.

players with cognitive levels n and n', and they play a Nash equilibrium of the game induced by their preferences. Given two preferences $u, u' \in U$, let $NE(u, u') \subseteq \Delta(A) \times \Delta(A)$ be the set of mixed equilibria of the game induced by the preferences u and u', i.e.

$$NE(u, u') = \{(\sigma, \sigma') \in \Delta(A) \times \Delta(A) : \sigma \in BR_u(\sigma') \text{ and } \sigma' \in BR_{u'}(\sigma)\}.$$

We are now in a position to define our key notion of a configuration (following the terminology of Dekel, Ely, and Yilankaya, 2007), by combining a type distribution with a behaviour policy, as represented by Nash equilibria and deception equilibria.

Definition 2. A *configuration* is a pair (μ, b) where $\mu \in \Delta(\Theta)$ is a type distribution, and $b = \left(b^N, b^D\right)$ is a *behaviour policy*, where $b^N, b^D : C(\mu) \times C(\mu) \longrightarrow \Delta(A)$ satisfy for each $\theta, \theta' \in C(\mu)$:

$$q(n_\theta, n_{\theta'}) + q(n_{\theta'}, n_\theta) < 1 \Rightarrow \left(b^N_\theta(\theta'), b^N_{\theta'}(\theta)\right) \in NE(\theta, \theta'), \text{ and}$$

$$q(n_\theta, n_{\theta'}) > 0 \Leftrightarrow n_\theta > n_{\theta'} \Rightarrow \left(b^D_\theta(\theta'), b^D_{\theta'}(\theta)\right) \in DE(\theta, \theta').$$

We interpret $b^D_\theta(\theta')$ (resp., $b^N_\theta(\theta')$) to be the strategy used by type θ against type θ' when deception occurs (resp., does not occur).

Given a configuration (μ, b) we call the types in the support of μ *incumbents*. Note that standard arguments imply that for any distribution μ, there exists a mapping $b : C(\mu) \times C(\mu) \to \Delta(A)$ such that (μ, b) is a configuration. Given a configuration (μ, b) and types $\theta, \theta' \in C(\mu)$, let $\pi_\theta(\theta' | (\mu, b))$ be the expected fitness of an agent with type θ conditional on being matched with θ':

$$\pi_\theta(\theta' | (\mu, b)) = (q(n_\theta, n_{\theta'}) + q(n_{\theta'}, n_\theta)) \cdot \pi\left(b^D_\theta(\theta'), b^D_{\theta'}(\theta)\right) + (1 - (q(n_\theta, n_{\theta'}) + q(n_{\theta'}, n_\theta))) \cdot \pi\left(b^N_\theta(\theta'), b^N_{\theta'}(\theta)\right).$$

The expected fitness of an individual of type θ in configuration (μ, b) is

$$\Pi_{\theta|(\mu,b)} = \sum_{\theta' \in C(\mu)} \mu(\theta') \cdot \pi_\theta(\theta' | (\mu, b)) - k_\theta,$$

where $\mu(\theta')$ denotes the frequency of type θ' in the population. Given a configuration (μ, b), let $\Pi_{(\mu,b)}$ be the average fitness in the population, i.e.,

$$\Pi_{(\mu,b)} = \sum_{\theta \in C(\mu)} \mu(\theta) \cdot \Pi_{\theta|(\mu,b)}.$$

When all incumbent types have the same expected fitness (i.e. $\Pi_{(\mu,b)} = \Pi_{\theta|(\mu,b)}$ for each $\theta \in C(\mu)$), we say that the configuration is *balanced*.

A number of aspects of our model of cognitive sophistication merit further discussion.

1. *Unidimensional cognitive ability*: In reality the ability to deceive and the ability to detect preferences are probably not identical. However, both of them are likely to be strongly related to cognitive ability in general, and more specifically to theory of mind and the ability to entertain higher-order intentional attitudes (Kinderman, Dunbar, and Bentall, 1998; Dunbar, 1998). For this reason we believe that a unidimensional cognitive trait is a reasonable approximation. Moreover, it is an approximation that affords us necessary tractability. We connect the abilities to detect and conceal preferences with the ability to deceive, by assuming (throughout the paper) that one is able to deceive one's opponent if and only if one observes the opponent's preferences and conceals one's own preferences from the opponent.

2. *Power of deception*: Our definition of deception equilibrium amounts to an assumption that a successful deception attempt allows the deceiver to implement her favourite strategy profile, under the constraint that the deceived party does not choose a dominated action from her point of view. Moreover, we assume that a player with a higher cognitive level knows whether her deception was successful when choosing her action. These assumptions give higher cognitive types a clear advantage over lower cognitive types. Hence, in an alternative model in which successful deceivers have less deception power, we would expect the evolutionary advantage of higher types to be weaker than in our current model. Below we find that (for generic games) in any stable state everyone plays the same efficient action profile and has the lowest cognitive level.[12] We conjecture that these states will remain stable also in a model where successful deception is less powerful. We leave for future research the analysis of feasible but less powerful deception technologies.

3. *Same deception against all lower types*: Our model assumes that a player may use different deceptions against different types with lower cognitive levels. We note that our results remain the same (with minor changes to the proofs) in an alternative setup in which individuals have to use the *same* mixed action in their deception efforts towards all opponents.

4. *Non-Bayesian deception*: Note that a successful deceiver is able to induce the opponent to believe that the deceiving type will play any mixed action $\hat{\sigma}$, even an action that is never played by any agent in the population. That is, deception is so powerful in our model that the deceived opponent is not able to apply Bayesian reasoning in his false assessment of which action the agent is going to play. We think of this assumption as describing a setting in which the deceiver (of a higher cognitive type) is able to provide a convincing argument (tell a convincing story) that she is going to play $\hat{\sigma}$. From a Bayesian perspective one might object that these arguments are signals that should be used to update beliefs. To this we would respond that the stories told to a potential victim by different deceivers will vary across

[12] Thus, in our setup a cognitive arms race (i.e. Machiavellian intelligence hypothesis à la Humphrey, 1976; Robson, 2003) is a non-equilibrium phenomenon, or alternatively a feature of non-generic games.

would-be deceivers, even across would-be deceivers with the same preferences. Hence no individual will ever accumulate a database containing more than one or a handful of similar arguments. The limited amount of data on similar arguments will preclude the efficient use of Bayesian updating for inferring likely behaviour following different arguments. We are not aware of the existence of a Bayesian model of deception that is satisfactory for our purposes. We leave the development of such a Bayesian model to future research.

5. *Observation and Nash equilibrium behaviour in the case of non-deception*: It is difficult to avoid an element of arbitrariness when making an assumption about what is being observed when neither party is able to deceive the other. As in most of the existing literature on the indirect evolutionary approach (e.g., Güth and Yaari, 1992; Dekel, Ely, and Yilankaya, 2007, Section 3), we assume that when there is no deception, then there is perfect observability of the opponent's preferences. In Section 4.1 we discuss the implications of the relaxation of this assumption. We consider it to be an important contribution of our analysis that it highlights the critical importance of the assumption made regarding observability, and the resulting behaviour, in matches without deception.
   We further assume that if two agents observe each other's preferences then they play a Nash equilibrium of the complete information game induced by their preferences. This assumption is founded on the common idea that when agents are not deceived, then (1) over time they adapt their beliefs (in a way that is consistent with Bayesian inference) about the distribution of actions they face, conditional on their partners' observed preferences, and (2) they best-reply given their belief about their current partner's distribution of actions. By contrast, as discussed above, when agents are deceived they are unable to correctly update their beliefs about their partner's action (i.e. unable to use Bayesian inference to arrive at beliefs about the opponent's distribution of actions). Still, they are able to best-reply given their (possibly false) beliefs about the deceiver's action.

6. *Continuum population and finite-support type distributions.* Our model is intended to be a simple approximation of a real-life environment that includes a large finite population, and in which new agents who join the population, or existing agents who revise their choice of type, typically choose to mimic one of the existing active types. As a result each active type is played by several agents (rather than by a single agent), and for each active type there is a positive probability of a match between agents who are endowed with this type. As is common in the literature, for tractability, we assume a continuum population and an "exact law of large numbers," rather then a large finite population. We want all other aspects of the model to be as close as possible to the real-life environment. Specifically, we want to maintain the property that for each type, there is a positive probability of a match between agents who are endowed with this type. In order to maintain this property, we have to

assume that the distribution of active types has a finite support.[13]

7. *Alternative interpretation of our model: social status.* As suggested by one of the referees, one can present an interesting interpretation of our model that describes social status, rather than deception. According to this interpretation, the level n_θ of type θ describes the social status (like caste) of agents belonging to this type. When two players are randomly matched to play a game, first a "social struggle" ensues. With a certain probability, the higher-caste player prevails and enslaves the lower-caste opponent. This means he can dictate the choice by the lower-caste opponent as long as the choice is undominated for this opponent. Otherwise, they simply play the Nash equilibrium of the game (given by their preferences). Maintaining a higher social status is costly in terms of fitness.

2.3 Evolutionary Stability

As discussed in the previous subsection, each agent in the population behaves in a way that maximises the agent's subjective preferences induced by the agent's type. By contrast, the distribution of types in the population evolves according to the expected material fitness obtained by each type. This evolutionary process is captured by the static solution concepts introduced in this subsection.

We consider dynamics in which types with higher expected fitness gradually become more frequent. One example of such dynamics is the replicator dynamic (Taylor and Jonker, 1978), which can be interpreted in terms of biological (asexual) reproduction or as social learning by imitation (see Weibull, 1995, Chapter 3 for a textbook introduction). According to the latter interpretation, an agent who has the opportunity to revise her choice or a new agent who joins the population randomly chooses a member of the population as "mentor," and imitates the mentor's type; the probability that an agent is chosen as a mentor is proportional to that agent's fitness.

Recall that a neutrally stable strategy (Maynard Smith and Price, 1973; Maynard Smith, 1982) is a strategy that, if played by most of the population, weakly outperforms any other strategy. Similarly, an evolutionarily stable strategy is a strategy that, if played by most of the population, strictly outperforms any other strategy.

Definition 3. A strategy $\sigma \in \Delta(A)$ is a *neutrally stable strategy (NSS)* if for every $\sigma' \in \Delta(A)$ there is some $\bar{\varepsilon} \in (0,1)$ such that if $\varepsilon \in (0,\bar{\varepsilon})$, then $\tilde{\pi}(\sigma', (1-\varepsilon)\sigma + \varepsilon\sigma') \leq \tilde{\pi}(\sigma, (1-\varepsilon)\sigma + \varepsilon\sigma')$. If weak inequality is replaced by strict inequality for each $\sigma' \neq \sigma$, then σ is an *evolutionarily stable strategy (ESS)*.

It is well known that NSSs and ESSs correspond to Lyapunov stable and asymptotically stable population states, respectively, under the replicator dynamics. That is, a population starting close

[13] More accurately, we need to assume that the set of active types is countable. All of our results hold under this somewhat weaker assumption.

to an NSS will always remain close to the NSS, and a population starting close to an ESS will converge to the ESS (see, e.g., Taylor and Jonker, 1978; Thomas, 1985; Bomze and Weibull, 1995; Cressman, 1997; Sandholm, 2010.)

We extend the notions of neutral and evolutionary stability from strategies to configurations. We begin by defining the type game that is induced by a configuration.

Definition 4. For any configuration (μ, b) the corresponding *type game* $\Gamma_{(C(\mu),b)}$ is the symmetric two-player game where each player's pure strategy space is $C(\mu)$, and the payoff to strategy θ, against θ', is $\pi_\theta(\theta' | (\mu, b)) - k_\theta$.

The definition of a type game allows us to apply notions and results from standard evolutionary game theory, where evolution acts upon strategies, to the present setting where evolution acts upon types. A similar methodology was used in Mohlin (2012). Note that each type distribution with support in $C(\mu)$ is represented by a mixed strategy in $\Gamma_{(C(\mu),b)}$.

We want to capture robustness with respect to small groups of individuals, henceforth called *mutants*, who introduce new types and new behaviours into the population. Suppose that a fraction ε of the population is replaced by mutants and suppose that the distribution of types within the group of mutants is $\mu' \in \Delta(\Theta)$. Consequently the post-entry type distribution is $\tilde{\mu} = (1-\varepsilon)\cdot\mu + \varepsilon\cdot\mu'$. That is, for each type $\theta \in C(\mu) \cup C(\mu')$, $\tilde{\mu}(\theta) = (1-\varepsilon)\cdot\mu(\theta) + \varepsilon\cdot\mu'(\theta)$. In line with most of the literature on the indirect evolutionary approach we assume that adjustment of behaviour is infinitely faster than adjustment of type distribution.[14] Thus we assume that the post-entry type distribution quickly stabilises into a configuration $\left(\tilde{\mu}, \tilde{b}\right)$. There may exist many such post-entry type configurations, all having the same type distribution, but different behaviour policies. We note that incumbents do not have to adjust their behaviour against other incumbents in order to continue playing Nash equilibria, and deception equilibria, among themselves. For this reason, we assume (similarly to Dekel, Ely, and Yilankaya, 2007, in the setup with perfect observability) that the incumbents maintain the same pre-entry behaviour among themselves. Formally:

Definition 5. Let (μ, b) and $\left(\tilde{\mu}, \tilde{b}\right)$ be two configurations such that $C(\mu) \subseteq C(\tilde{\mu})$. We say that $\left(\tilde{\mu}, \tilde{b}\right)$ is *focal* (with respect to (μ, b)) if $\theta, \theta' \in C(\mu)$ implies that $\tilde{b}^D_\theta(\theta') = b^D_\theta(\theta')$ and $\tilde{b}^N_\theta(\theta') = b^N_\theta(\theta')$.

Standard fixed-point arguments imply that for every configuration (μ, b) and every type distribution $\tilde{\mu}$ satisfying $C(\mu) \subseteq C(\tilde{\mu})$, there exists a behaviour policy $\tilde{b}$ such that $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration.

Our stability notion requires that the incumbents outperform all mutants in all configurations that are focal relative to the initial configuration.

[14] Sandholm (2001) and Mohlin (2010) are exceptions.

Definition 6. A configuration (μ, b) is a *neutrally stable configuration (NSC)* if, for every $\mu' \in \Delta(\Theta)$, there is some $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$, it holds that if $\left(\tilde{\mu}, \tilde{b}\right)$, where $\tilde{\mu} = (1-\varepsilon) \cdot \mu + \varepsilon \cdot \mu'$, is a focal configuration, then μ is an NSS in the type game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. The configuration (μ, b) is an *evolutionarily stable configuration (ESC)* if the same conditions imply that μ is an ESS in the type game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$ for each $\mu' \neq \mu$.

We conclude this section by discussing a few issues related to our notion of stability.

1. In line with existing notions of evolutionary stability in the literature (in particular, the notions of Dekel, Ely, and Yilankaya, 2007, and Alger and Weibull, 2013), we require the mutants to be outperformed in all focal configurations (rather than requiring them to be outperformed in at least one focal configuration). This reflects the assumption that the population converges to a new post-entry equilibrium in a decentralised (possibly random) way that may lead to any of the post-entry focal configurations. Thus the incumbents cannot coordinate their post-entry play on a specific focal configuration that favors them.

2. In order to be consistent with the standard definition of neutral stability, we require the incumbents to earn weakly more than the average payoff of the mutants. We note that all of our results remain the same if one uses an alternative weaker definition that requires the incumbents to earn weakly more than the worst-performing mutant.

3. The main stability notion that we use in the paper is NSC. The stronger notion of ESC is not useful in our main model because there always exist equivalent types that have slightly different preferences (as the set of preferences is a continuum) and induce the same behaviour as the incumbents. Such mutants always achieve the same fitness as the incumbents in post-entry configurations, and thus ESCs never exist. Note that the stability notions in Dekel, Ely, and Yilankaya (2007) and Alger and Weibull (2013) are also based on neutral stability.[15] In Section B we study a variant of the model in which the preferences may depend also on the opponent's types. This allows for the existence of ESCs.

4. Observe that Definition 6 implies internal stability with respect to small perturbations in the frequencies of the incumbent types (because when $\mu' = \mu$, then μ is required to be an NSS in $\Gamma_{(C(\mu),b)}$). By standard arguments, internal stability implies that any NSC is balanced: all incumbent types obtain the same fitness.

5. The stability notions of Dekel, Ely, and Yilankaya (2007) and Alger and Weibull (2013) consider only monomorphic groups of mutants (i.e. mutants all having the same type). We additionally consider stability against polymorphic groups of mutants (as do Herold and

[15] In their stability analysis of *homo hamiltonensis* preferences Alger and Weibull (2013) disregard mutants who are behaviourally indistinguishable from *homo hamiltonensis* upon entry.

Kuzmics, 2009). One advantage of our approach is that it allows us to use an adaptation of the well-known notion of ESS, which immediately implies dynamic stability and internal stability, whereas Dekel, Ely, and Yilankaya (2007) have to introduce a novel notion of stability without these properties. Remark 3 below discusses the influence on our results of using an alternative definition that deals only with monomorphic mutants.

3 Results

3.1 Preliminary Definitions

Define the *deviation gain* of action $a \in A$, denoted by $g(a) \in \mathbb{R}^+$, as the maximal gain a player can get by playing a different action in a population in which everyone plays a:

$$g(a) = \max_{a' \in A} \pi(a', a) - \pi(a, a).$$

Note that $g(a) = 0$ iff (a, a) is a Nash equilibrium.

Define the *effective cost of deception* in the environment, denoted by $c \in \mathbb{R}^+$, as the minimal ratio between the cognitive cost and the probability of deceiving an opponent of cognitive level one:[16] [17]

$$c = \min_{n \geq 2} \frac{k_n}{q(n, 1)}.$$

We say that a strategy profile is efficient if it maximises the sum of fitness payoffs. Formally:

Definition 7. A strategy profile (σ, σ') is *efficient* in the game $G = (A, \pi)$ if $\pi(\sigma, \sigma') + \pi(\sigma', \sigma) \geq \pi(a, a') + \pi(a', a)$, for each action profile $(a, a') \in A^2$.

Note that our notion of efficiency is defined: (1) with respect to the fitness payoff (rather than the agents' subjective payoffs), similarly to the analogous definition of efficiency in Dekel, Ely, and Yilankaya (2007), and (2) with respect to the strategy profile played by the agents; by contrast, the definition does not take into account the cognitive costs.

A pure Nash equilibrium (a, a) is *strict* if $\pi(a, a) > \pi(a', a)$ for all $a' \neq a \in A$. Let $\hat{\pi} = \max_{a,a' \in A} (0.5 \cdot (\pi(a, a') + \pi(a', a)))$ denote the efficient payoff, i.e. the average payoff achieved by players who play an efficient profile.

[16] The minimum in the definition of c is well defined for the following reason. Let $\hat{n}$ be a number such that $k_{\hat{n}} > \frac{k_2}{q(2,1)}$ (such a number exists because $\lim_{n \to \infty} k_n = \infty$). Observe that $\frac{k_n}{q(n,1)} \geq k_n > \frac{k_2}{q(2,1)}$ for any $n \geq \hat{n}$. This implies that there is an $\bar{n}$ such that $2 \leq \bar{n} \leq \hat{n}$ and $\bar{n} = \arg\min_{n \geq 2} \frac{k_n}{q(n,1)}$.

[17] We define the effective cost of deception only with respect to an opponent with a cognitive level of one because we later show (Lemma 1 and Theorem 2) that the only candidate to be an NSC is a configuration in which all agents have a cognitive level of one, and such a configuration is an NSC iff the effective cost of defection against these incumbents with $n = 1$ is sufficiently large.

An action a is a *punishment action* if playing it guarantees that the opponent will obtain less than the efficient payoff, i.e. $\pi(a', a) < \hat{\pi}$ for each $a' \in A$. Some of our results below assume that the underlying game admits a punishment action.

Remark 2. Many economic interactions admit punishment actions. A few examples include:

1. Price competition (Bertrand), either for a homogeneous good or for differentiated goods, where a punishment action is setting the price equal to zero.
2. Quantity competition (Cournot), either for a homogeneous good or for differentiated goods, where the punishment action is "flooding" the market.
3. Public good games, where contributing nothing to the public good is the punishment action.
4. Bargaining situations, where the punishment action is for one side of the bargaining to insist on obtaining all surplus.
5. Any game that admits an action profile that Pareto dominates all other action profiles (i.e., games with common interests).

Moreover, if one adds to any underlying generic game a new pure action that is equivalent to playing the mixed action that min-maxes the opponent's payoff (e.g., in matching pennies this new action is equivalent to privately tossing a coin and then playing according to the toss's outcome), then this newly added action is always a punishment action.

Given a configuration (μ, b) let $\bar{n} = \max_{\theta \in C(\mu)} n_\theta$ denote the maximal cognitive level of the incumbents. We refer to incumbents with this cognitive level as the *highest types*.

A deception equilibrium is *fitness maximising* if it maximises the fitness of the higher type in the match (under the restriction that the lower type plays an action that is not dominated, given her preferences). Formally:

Definition 8. Let θ, θ' be types with $n_\theta > n_{\theta'}$. A deception equilibrium $(\tilde{\sigma}, \tilde{\sigma}')$ is *fitness maximising* if

$$(\tilde{\sigma}, \tilde{\sigma}') \in \arg \max_{\sigma \in \Delta(A),\, \sigma' \in \Sigma(u_{\theta'})} \pi(\sigma, \sigma').$$

Let $FMDE(\theta, \theta') \subseteq DE(\theta, \theta')$ denote the set of all such fitness-maximising deception equilibria of two types θ, θ' with $n_\theta > n_{\theta'}$. In principle, $FMDE(\theta, \theta')$ might be an empty set (if there is no action profile that maximises both the fitness and the subjective utility of the higher type). Our first result (Theorem 1 below) implies that the preference of the higher type in any NSC are such that the set $FMDE(\theta, \theta')$ is non-empty.

A configuration is pure if everyone plays the same action. Formally:

Definition 9. A configuration (μ, b) is *pure* if there exists $a^* \in A$ such that for each $\theta, \theta' \in C(\mu)$ it holds that $b_\theta^N(\theta') = a^*$ whenever $q(\theta, \theta') < 1$, and $b_\theta^D(\theta') = a^*$ whenever $q(\theta, \theta') > 0$. With a slight abuse of notation we denote such a pure configuration by (μ, a^*), and we refer to $b \equiv a^*$ as the *outcome* of the configuration.

In order to simplify the notation and the arguments in the proofs, we assume throughout this section that the underlying game admits at least three actions (i.e. $|A| \geq 3$). If the original game has only two actions, then adding a third action, which is dominated by the other two actions, allows all the arguments in the proof to work. More complicated (and less instructive) variants of the proofs can also be applied to a game with two actions without adding a third, dominated action.

3.2 Characterisation of the Highest Types' Behaviour

In this section we characterise the behaviour of an incumbent type, $\bar{\theta} = (u, \bar{n})$, which has the highest level of cognition in the population.[18] We show that the behaviour satisfies the following three conditions:

1. Type $\bar{\theta}$ plays an efficient action profile when meeting itself.
2. Type $\bar{\theta}$ maximises its fitness in all deception equilibria.
3. Any opponent with a lower cognitive level achieves at most the efficient payoff $\hat{\pi}$ against type $\bar{\theta}$.

Theorem 1. *Let (μ^*, b^*) be an NSC, and let $\underline{\theta}, \bar{\theta} \in C(\mu^*)$. Then: (1) if $n_{\bar{\theta}} = \bar{n}$ then $\pi\left(\bar{\theta}, \bar{\theta}\right) = \hat{\pi}$; (2) if $n_{\underline{\theta}} < n_{\bar{\theta}} = \bar{n}$ then $\left(b_{\bar{\theta}}^D\left(\underline{\theta}\right), b_{\underline{\theta}}^D\left(\bar{\theta}\right)\right) \in FMDE\left(\bar{\theta}, \underline{\theta}\right)$; and (3) if $n_{\underline{\theta}} < n_{\bar{\theta}} = \bar{n}$ then $\pi\left(\underline{\theta}, \bar{\theta}\right) \leq \hat{\pi}$.*

Proof Sketch (formal proof in Appendix A.2). The proof utilises mutants (denoted by $\theta_1, \theta_2, \theta_3$, and $\hat{\theta}$, below) with the highest cognitive level $\bar{n}$ and with a specific kind of utility function, called *indifferent and pro-generous*, that makes a player indifferent between all her own actions, but which makes the player prefer that the opponent choose an action that allows the player to obtain the highest possible fitness payoff.

To prove part 1 of the theorem, assume to the contrary that $\pi\left(b_{\bar{\theta}}\left(\bar{\theta}\right), b_{\bar{\theta}}\left(\bar{\theta}\right)\right) < \hat{\pi}$. Let $a_1, a_2 \in A$ be any two actions such that (a_1, a_2) is an efficient action profile (i.e. $0.5 \cdot (\pi(a_1, a_2) + \pi(a_1, a_2)) = \hat{\pi}$). Consider three different mutant types θ_1, θ_2, and θ_3, which are of the highest cognitive level that is present in the population, and have indifferent and pro-generous utility functions. Suppose

[18] For tractability we assume that a configuration can have only finite support. Note, however, that there is some sufficiently high cognitive level n such that $k_n > \max_{a,a' \in A} \pi(a, a')$. As a result, even if one relaxes the assumption of finite support, any NSC must include only a finite number of cognitive levels, also without the finite-support assumption.

equal fractions of these three mutant types enter the population.[19] There is a focal post-entry configuration in which the incumbents keep playing their pre-entry play among themselves, the mutants play the same Nash equilibria as the incumbent $\bar{\theta}$ against all incumbent types (and the incumbents behave against the mutants in the same way they behave against $\bar{\theta}$), the mutants play fitness-maximising deception equilibria against all lower types, when mutants of type θ_i are matched with mutants of type[20] $\theta_{(i+1) \bmod 3}$ they play the efficient profile (a_1, a_2), and when two mutants of the same type are matched they play the same way as two incumbents of type $\bar{\theta}$ that are matched together. In such a focal post-entry configuration all mutants earn a weakly higher fitness than $\bar{\theta}$ against the incumbents, and a strictly higher fitness against the mutants. This implies that (μ^*, b^*) cannot be an NSC.

To prove part 2, assume to the contrary that $\left(b^D_{\bar{\theta}}(\underline{\theta}), b^D_{\underline{\theta}}\left(\bar{\theta}\right)\right) \notin FMDE\left(\bar{\theta}, \underline{\theta}\right)$. Suppose mutants of type $\hat{\theta}$ enter. Consider a post-entry configuration in which the incumbents keep playing their pre-entry play among themselves, and the mutants mimic the play of $\bar{\theta}$, except that they play fitness-maximising deception equilibria against all lower types. The mutants obtain a weakly higher payoff than $\bar{\theta}$ against all types, and a strictly higher payoff than $\bar{\theta}$ against at least one lower type. Thus (μ^*, b^*) cannot be an NSC.

To prove part 3, assume to the contrary that $\pi\left(\underline{\theta}, \bar{\theta}\right) > \hat{\pi}$. This implies that against type $\bar{\theta}$, type $\underline{\theta}$ earns more than $\hat{\pi}$ in either the deception equilibrium or the Nash equilibrium. Suppose mutants of type $\hat{\theta}$ enter. Consider a post-entry configuration in which the incumbents keep playing their pre-entry play among themselves, while the mutants: (i) play fitness-maximising deception equilibria against lower types, (ii) mimic type $\underline{\theta}$'s play in the Nash/deception equilibrium against type $\bar{\theta}$ in which $\underline{\theta}$ earns more than $\hat{\pi}$, and (iii) mimic the play of $\bar{\theta}$ in all other interactions. The type $\hat{\theta}$ mutants earn strictly more than $\bar{\theta}$ against both $\hat{\theta}$ and $\bar{\theta}$. The mutants earn weakly more than $\bar{\theta}$ against all other types. This implies that (μ^*, b^*) cannot be an NSC. □

Remark 3. The first part of Theorem 1 (a highest type must play an efficient strategy profile when meeting itself) is similar to Dekel, Ely, and Yilankaya's (2007) Proposition 2, which shows that only efficient outcomes can be stable in a setup with perfect observability and no deception. We should note that Dekel, Ely, and Yilankaya (2007) use a weaker notion of efficiency. An action is efficient in the sense of Dekel, Ely, and Yilankaya (2007) (DEY-efficient) if its fitness is the highest among the symmetric strategy profiles (i.e. action a is DEY-efficient if $\pi(a, a) \geq \pi(\sigma, \sigma)$ for all strategies $\sigma \in \Delta(A)$). Observe that our notion of efficiency (Definition 7) implies DEY-efficiency, but the converse is not necessarily true. The weaker notion of DEY-efficiency is the relevant one in the setup of Dekel, Ely, and Yilankaya (2007), because they consider only monomorphic groups

[19] One must have at least two different types of mutants, in order for the mutants to be able to play the asymmetric profile (a_1, a_2). We preset a construction with three different mutant types in order to allow all mutant types to outperform the incumbents (one can also prove the result using a constructions with only two different mutant types, but in this case one can only guarantee that the mutants, on average, would outperform the incumbents)

[20] If $i = 1$ (resp., $i = 2$, $i = 3$), then $\theta_{(i+1) \bmod 3} = \theta_2$ (resp., $\theta_{(i+1) \bmod 3} = \theta_3$, $\theta_{(i+1) \bmod 3} = \theta_1$).

mutants; i.e. all mutants who enter at the same time are of the same type. A similar result would also hold in our setup if we imposed a similar limitation on the set of feasible mutants. However, without such a limitation, heterogeneous mutants can correlate their play, and our stronger notion of efficiency is required to characterise stability.

An immediate corollary of Theorem 1 is that a game that has only asymmetric efficient action profiles does not admit any NSCs.

Corollary 1. *If G does not have an efficient profile that is symmetric (i.e. if $\pi(a,a) < \hat{\pi}$ for each $a \in A$), then the game does not admit an NSC.*

Remark 4. As discussed in Remark 1, any interaction (symmetric or asymmetric) can be embedded in a larger, symmetric game in which nature first randomly assigns roles to the players, and then each player chooses an action given his assigned role.[21] Observe that *such an embedded game always admits an efficient symmetric action profile*. In particular, if the efficient asymmetric profile in the original game is (a, a'), then the efficient symmetric profile in the embedded game is the one in which each player plays a as the row player and a' as the column player.

3.3 Characterisation of Pure NSCs

In this subsection we characterise pure NSCs, i.e. stable configurations in which everyone plays the same pure action in every match. Such a configuration may be viewed as representing the state of a population that has settled on a convention that there is a unique correct way to behave. We begin by showing that in a pure NSC all incumbents have the minimal cognitive level, since having a higher ability does not yield any advantage when everyone plays the same action.

Lemma 1. *If (μ, a^*) is an NSC, and $(u,n) \in C(\mu)$, then $n = 1$.*

Proof. Since all players earn the same game payoff of $\pi(a^*, a^*)$, they must also incur the same cognitive cost, or else the fitness of the different incumbent types would not be balanced (which would contradict that (μ, a^*) is an NSC). Moreover, this uniform cognitive level must be level 1. Otherwise a mutant of a lower level, who strictly prefers to play a^* against all actions, would strictly outperform the incumbents in nearby post-entry focal configurations. □

The following proposition shows that a pure outcome is stable iff it is efficient and its deviation gain is smaller than than the effective cost of deception. Formally:

Proposition 1. *Let a^* be an action in a game that admits a punishment action. The following two statements are equivalent:*

(a) There exists a type distribution μ such that (μ, a^*) is an NSC.

(b) a^* satisfies the following two conditions: (1) $\pi(a^*, a^*) = \hat{\pi}$, and (2) $g(a^*) \leq c$.

[21] If the original game is symmetric, the role (i.e. being either the row or the column player) can be interpreted as reflecting some observable payoff-irrelevant asymmetry between the two players.

Proof.

1. *"If side."* Assume that (a^*, a^*) is an efficient profile and that $g(a^*) \leq c$. Let $\widetilde{a}$ be a punishment action. Consider a monomorphic configuration (μ, a^*) consisting of type $\theta^* = (u^*, 1)$ where all incumbents are of cognitive level 1 and of the same preference type u^*, according to which all actions except a^* and $\widetilde{a}$ are strictly dominated, $\widetilde{a}$ weakly dominates a^*, and a^* is a best reply to itself:

$$u^*(a, a') = \begin{cases} 1 & \text{if } a = \widetilde{a} \text{ and } a' \neq a^* \\ 0 & \text{if } a = a^* \text{ or } (a = \widetilde{a} \text{ and } a' = a^*.) \\ -1 & \text{otherwise.} \end{cases}$$

Consider first mutants with cognitive level one. Observe that in any post-entry configuration mutants with cognitive level one earns at most $\hat{\pi}$ when they are matched with the incumbents, and strictly less than $\hat{\pi}$ if the mutants play any action $a \neq a^*$ with positive probability against the incumbents. Further observe, that the mutants can earn (on average) at most $\hat{\pi}$ when they are matched with other mutants (because $\hat{\pi}$ is the efficient payoff). This implies that incumbents weakly outperform any mutants with cognitive level one in any post-entry population.

Next, consider mutants with a higher cognitive level $n > 1$. Such mutants can earn at most $\hat{\pi} + g(a^*)$ when they deceive the incumbents and at most $\hat{\pi}$ when they do not deceive the incumbents (recall that $\pi(\widetilde{a}, \widetilde{a}) + g(\widetilde{a}) = max_{a'} \pi(a', \widetilde{a}) < \hat{\pi}$ because $\widetilde{a}$ is a punishment action). Thus the mutants are weakly outperformed by the incumbents if

$$q(n,1) \cdot (g(a^*) + \hat{\pi}) + (1 - q(n,1)) \cdot \hat{\pi} - k_n \leq \hat{\pi} \;\Leftrightarrow\; g(a^*) \leq \frac{k_n}{q(n,1)}.$$

This holds for all n if $g(a^*) \leq c$. Thus, the probability of deceiving the incumbents is at most $\frac{k_n}{g(a^*)}$. The fact that $g(a^*) \leq c$ implies that the average payoff of the mutants against the incumbents is less than $\hat{\pi} + g(a^*) \cdot \frac{k_n}{g(a^*)} \leq \hat{\pi} + k_n$, and thus if the mutants are sufficiently rare they are weakly outperformed (due to paying the cognitive cost of k_n). We conclude that (μ, a^*) is an NSC.

2. *"Only if side."* Assume that (μ, a^*) is an NSC. Theorem 1 implies that $\pi(a^*, a^*) = \hat{\pi}$. Assume that $g(a^*) > c$. The definition of the effective cost of deception implies that there exists a cognitive level n such that $\frac{k_n}{q(n,1)} < g(a^*)$. Lemma 1 implies that all the incumbents have cognitive level 1. Consider mutants with cognitive level n and completely indifferent preferences (i.e. preferences that induce indifference between all action profiles). Let a' be a best reply against a^*. There is a post-entry focal configuration in which (i) the incumbents play a^* against the mutants, (ii) the mutants play a' when they deceive an incumbent

opponent and a^* when they do not deceive an incumbent opponent, and (iii) the mutants play a^* when they are matched with another mutant. Note that the mutants achieve at least $\hat{\pi} + g(a^*) \cdot q(n,1)$ when they are matched against the incumbents. The gain relative to incumbents, $g(a^*) \cdot q(n,1)$, outweighs their additional cognitive cost of k_n, by our assumption that $g(a^*) > c$. Thus the mutants strictly outperform the incumbents.

□

3.4 Characterisation of NSCs in Generic Games

In this section we characterise NSCs in generic games, by which we mean games in which any two different action profiles each give the same player a different payoff, and each yield a different sum of payoffs.

Definition 10. A (symmetric) game is generic if for each $a, a', b, b' \in A$, $\{a, a'\} \neq \{b, b'\}$ implies

$$\pi(a, a') \neq \pi(b, b'), \text{ and } \pi(a, a') + \pi(a', a) \neq \pi(b, b') + \pi(b', b).$$

For example, if the entries of the payoff matrix π are drawn independently from a continuous distribution on an open subset of the real numbers, then the induced game is generic with probability one.

Note that a generic game admits at most one efficient action profile. From Corollary 1 we know that if the game does not have a symmetric efficient profile then it does not admit any NSC (and as discussed in Remark 4, essentially every interaction admits a symmetric efficient profile). Hence we can restrict attention to games with exactly one efficient action profile. Let $\bar{a}$ denote the action played in this unique profile.

Next we present our main result: all incumbent types play efficiently in any NSC of a generic game.

Theorem 2. *If (μ^*, b^*) is an NSC of a generic game with a (unique) efficient outcome $(\bar{a}, \bar{a})$, then $b^* \equiv \bar{a}$, for all $\theta, \theta' \in C(\mu^*)$; i.e. all types play the pure action $\bar{a}$ in all matches.*

Proof. Assume to the contrary that configuration (μ^*, b^*) is an NSC such that there are some $\theta, \theta' \in C(\mu^*)$ such that $b_\theta^N(\theta') \neq \bar{a}$ and $q(\theta_n, \theta_{n'}) + q(\theta_{n'}, \theta_n) < 1$, or $b_\theta^D(\theta') \neq \bar{a}$ and $q(\theta_n, \theta_{n'}) > 0$. Let $\mathring{\theta}$ be the type with the highest cognitive level among the types that satisfy at least one of the following conditions:

(A) $\mathring{\theta}$ plays inefficiently against itself, i.e. $\pi\left(\mathring{\theta}, \mathring{\theta}\right) < \hat{\pi}$.

(B) $\mathring{\theta}$ and an opponent with a weakly higher type play an inefficient strategy profile, i.e. $0.5 \cdot \left(\pi\left(\mathring{\theta}, \theta'\right) + \pi\left(\theta', \mathring{\theta}\right)\right) < \hat{\pi}$ for some $\theta' \neq \mathring{\theta}$ with $n_{\mathring{\theta}} \leq n_{\theta'}$.

(C) A strictly lower type earns strictly more than $\hat{\pi}$ against $\mathring{\theta}$, i.e. $\pi\left(\theta'', \mathring{\theta}\right) > \hat{\pi}$ for some $\theta'' \neq \mathring{\theta}$ with $n_{\mathring{\theta}} > n_{\theta''}$.

We will now successively rule out each of these cases.

Assume first that (A) holds. Let $\hat{u}$ be a utility function that is identical to $u_{\mathring{\theta}}$ except that: (i) the payoff of the outcome $(\bar{a}, \bar{a})$ is increased by the minimal amount required to make it a best reply to itself, and (ii) the payoff of some other outcome is altered slightly (to ensure $\hat{u}$ is not already an incumbent) in a way that does not change the behaviour of agents. (The formal definition of $\hat{u}$ is provided in Appendix A.3.) Suppose that mutants of type $\hat{\theta} = (\hat{u}, n_\theta)$ invade the population. Consider a focal post-entry configuration in which the mutants mimic the play of the type $\mathring{\theta}$ incumbents in all matches except that: (i) the mutants play the efficient profile $(\bar{a}, \bar{a})$ among themselves (which yields a higher payoff than what $\bar{\theta}$ achieves when matched against $\mathring{\theta}$), and (ii) when the mutants face a higher type they play either $(\bar{a}, \bar{a})$ or the same deception/Nash equilibrium that the higher types play against $\bar{\theta}$. It follows that the mutants $\hat{\theta}$ earn a strictly higher payoff than $\mathring{\theta}$ against $\hat{\theta}$, and a weakly higher fitness than type θ against all other types. Thus the mutants strictly outperform the incumbents, which contradicts the assumption that (μ^*, b^*) is an NSC. The full technical details of this argument are given in Appendix A.3.

Next, assume that case (B) holds and that case (A) does not hold. This implies that

$$0.5 \cdot \left(\pi\left(\mathring{\theta}, \theta'\right) + \pi\left(\theta', \mathring{\theta}\right)\right) < \hat{\pi} = \pi\left(\mathring{\theta}, \mathring{\theta}\right) = \pi\left(\theta', \theta'\right).$$

That is, in the subpopulation that includes types $\mathring{\theta}$ and θ' the within-type matchings yield higher payoffs than out-group matchings (an average payoff of less than $\hat{\pi}$). The following formal argument shows that this property implies dynamic instability. The fact that (μ^*, b^*) is an NSC implies that μ^* is an NSS in the type game $\Gamma_{(\mu^*, b^*)}$. Let $\mathbf{B}$ be the payoff matrix of the type game $\Gamma_{(\mu^*, b^*)}$ and let $n = |C(\mu^*)|$. It is well known (e.g., Hofbauer and Sigmund, 1988, Exercise 6.4.3, and Hofbauer, 2011, pp. 1–2) that an interior Nash equilibrium of a normal-form game is an NSS if and only if the payoff matrix is negative semi-definite with respect to the tangent space, i.e. if and only if $x^T \mathbf{B} x \leq 0$ for each $x \in \mathbb{R}^n$ such that $\sum_i x_i = 0$. Assume without loss of generality that type $\mathring{\theta}$ (θ') is represented by the j^{th} (k^{th}) row of the matrix B. Let the column vector x be defined as follows: $x(j) = 1$, $x(k) = -1$, and $x(i) = 0$ for each $i \notin \{j, k\}$. That is, the vector x has all entries equal to zero, except for the j^{th} entry, which is equal to 1, and the k^{th} entry, which is equal to -1. We have

$$\begin{aligned}
x^T \mathbf{B} x &= B_{jj} - B_{kj} - B_{jk} + B_{kk} \\
&= \pi(\bar{a}, \bar{a}) - k_{n_{\mathring{\theta}}} + \pi(\bar{a}, \bar{a}) - k_{n_{\theta'}} - \left(\pi\left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right) - k_{n_{\mathring{\theta}}} + \pi\left(b_{\theta'}\left(\mathring{\theta}\right), b_{\mathring{\theta}}(\theta')\right) - k_{n_{\theta'}}\right) \\
&= 2 \cdot \pi(\bar{a}, \bar{a}) - \left(\pi\left(b_{\mathring{\theta}}(\theta'), b_{\theta'}\left(\mathring{\theta}\right)\right) + \pi\left(b_{\theta'}\left(\mathring{\theta}\right), b_{\mathring{\theta}}(\theta')\right)\right) > 0.
\end{aligned}$$

Thus $\mathbf{B}$ is not negative semi-definite.

Finally, assume that only case (C) holds. Let $\bar{\theta}$ be an incumbent type with the highest cognitive level. The fact that case (B) does not hold implies that $\pi\left(\bar{\theta},\mathring{\theta}\right)=\pi\left(\mathring{\theta},\bar{\theta}\right)=\hat{\pi}$. The fact that case (C) holds implies that $\pi\left(\theta'',\mathring{\theta}\right)>\hat{\pi}$, which implies that type $\mathring{\theta}$ has an undominated action that can yield a deceiving opponent a payoff of more than $\hat{\pi}$ in a deception equilibrium. This contradicts part (2) of Theorem 1, according to which we should have $\left(b_{\bar{\theta}}^{D}\left(\mathring{\theta}\right),b_{\mathring{\theta}}^{D}\left(\bar{\theta}\right)\right)=FMDE\left(\bar{\theta},\mathring{\theta}\right)$. We have shown that no type in the population satisfies either (A), (B), or (C). The fact that no type satisfies (A) implies that in any match of agents of the same type both agents play action $\bar{a}$, and the fact that no type satisfies (B) implies that in any match between two agents of different types both players play action $\bar{a}$. □

Combining the results of this section with the above characterisation of pure NSCs yields the following corollary, which fully characterises the NSCs of generic games that admit punishment actions (as discussed in Remark 2, such actions exist in many economic interactions). The result shows that such games admit an NSC iff the deviation gain from the pure efficient symmetric profile is smaller than the effective cost of defection, and when an NSC exists, its outcome is the pure efficient symmetric profile. In particular, in any game that admits an efficient symmetric pure Nash equilibrium, this equilibrium is the unique NSC outcome, and in the Prisoner's Dilemma mutual cooperation is the unique NSC outcome iff the gain from defecting against a cooperator is less than the effective cost of deception, and no NSC exists otherwise.

Corollary 2. *Let G be a generic game that admits a punishment action. The environment admits an NSC iff there exists an efficient symmetric pure profile (a^*,a^*) satisfying $g(a^*)\leq c$ (i.e. the deviation gain is smaller than the effective cost of deception). Moreover, if (μ,b) is an NSC, then $b\equiv a^*$, and $n=1$ for all $(u,n)\in C(\mu)$.*

Remark 5. Corollary 2 shows that generic games do not admit NSCs if the effective cost of deception is less than the deviation gain of the efficient profile. In such cases the distribution of types and their induced behaviour will not converge to a static population state. We leave the formal analysis of environments that do not admit NSCs to future research. One conjecture for the dynamic behaviour in such environments is a never-ending cycle between states in which almost all agents are naive and play an efficient action profile, and states in which different cognitive levels coexist, and agents play inefficient action profiles (see the related analysis of cyclic behaviour in the Prisoner's Dilemma with cheap talk and material preferences in Wiseman and Yilankaya, 2001).

Remark 6. Corollary 2 states that in an NSC of a generic game everyone has the same cognitive level. One may wonder how this relates to the apparent cognitive heterogeneity in the real world. Our analysis in this paper assumes a single underlying game, while in reality we face a potentially infinite set of games. If an individual's fitness is the result of interactions in a set of games that includes generic games with an NSC as well as non-generic games (see Section 3.5) or generic

games that do not admit any NSC (see previous remark), then evolution may lead to states in which different cognitive levels coexist, possibly with a never-ending cycle between states with different mixtures of cognitive levels.

Remark 7. Corollary 2 assumes that the underlying game admits a punishment action $\tilde{a}$, that gives an opponent a payoff strictly smaller than the efficient payoff $\hat{\pi}$, regardless of the opponent's play. This punishment action is used in the construction of the NSC that induces the efficient action a^*. Specifically, a non-deceived incumbent plays the punishment action a' against any mutant who does not always plays action a^*. If the game does not admit a punishment action, then (1) a complicated game-specific construction of the way in which incumbents behave against mutants who do not always play a^* may be required to support the efficient action as the outcome of an NSC, and (2) this construction may require further restrictions on the effective cost of deception, in addition to $g(a^*) \leq c$. We leave the study of these issues to future research.

3.5 Non-Pure NSCs in Non-generic Games

The previous two subsections fully characterise (i) pure NSCs and (ii) NSCs in generic games. In this section we analyse non-pure NSCs in non-generic games. Non-generic games may be of interest in various setups, such as: (1) normal-form representation of generic extensive-form games (the induced matrix is typically non-generic), and (2) interesting families of games, such as zero-sum games. Unlike generic games, non-generic games can admit NSCs that are not pure and that may therefore contain multiple cognitive levels. To demonstrate this we consider the Rock-Paper-Scissors game, with the following payoff matrix:[22]

$$\begin{array}{c|ccc} & R & P & S \\ R & 0,0 & -1,1 & 1,-1 \\ P & 1,-1 & 0,0 & -1,1 \\ S & -1,1 & 1,-1 & 0,0 \end{array}.$$

To simplify the analysis and the notations we assume in this subsection that a player always succeeds in deceiving an opponent with a lower cognitive level, i.e. that $q(n,n') = 1$ whenever $n > n'$. The analysis can be extended to the more general setup.

The result below shows that, under mild assumptions on the cognitive cost function, this game admits an NSC in which all players have the same materialistic preferences, but players of different cognitive levels coexist, and non-Nash profiles are played in all matches between two individuals

[22]For the construction presented in this subsection to work, the underlying game must be non-generic. Observe that if one slightly perturbs the payoffs of the Rock-Paper-Scissors game to make it a strictly competitive almost-zero-sum generic game, then Corollary 2 applies, and the only candidate to be an NSC is a configuration in which all agents have cognitive level one, and they all play an efficient action profile.

of different cognitive levels. More precisely, when individuals of different cognitive levels meet, the higher-level individual deceives the lower-level individual into taking a pure action that the higher-level individual then best-replies to. Thus the higher-level individual earns 1 and her opponent earns -1. Individuals of the same cognitive level play the unique Nash equilibrium. This means that higher-level types will obtain a payoff of 1 more often than lower-level types, and lower-level types will obtain a payoff of -1 more often than higher-level types. In the NSC this payoff difference is offset exactly by the higher cognitive cost paid by the higher types. Moreover, the cognitive cost is increasing and unbounded such that at some point the cost of cognition outweighs any payoff differences that may arise from the underlying game. This implies that there is an upper bound on the cognitive sophistication in the population.

Proposition 2. *Let G be a Rock-Paper-Scissors game. Let u^{π} denote the (materialistic) preference such that $u^{\pi}(a, a') = \pi(a, a')$ for all profiles (a, a'). Assume that $q(n, n') = 1$ whenever $n \neq n'$. Further assume that the marginal cognitive cost is small but non-vanishing, so that (a) there is an N such that $k_N \leq 2 < k_{N+1}$, and (b) it holds that $1 > k_{n+1} - k_n$ for all $n \leq N$. Under these assumptions there exists an NSC (μ^*, b^*) such that $C(\mu^*) \subseteq \{(u^{\pi}, n)\}_{n=1}^{N}$, and μ^* is mixed (i.e. $|C(\mu^*)| > 1$). The behaviour of the incumbent types is as follows: if the individuals in a match are of different cognitive levels, then the higher level plays Paper and the lower level plays Rock; if both individuals in a match are of the same cognitive level, then they both play the unique Nash equilibrium (i.e. randomise uniformly over the three actions).*

Appendix C contains a formal proof of this result and relates it to a similar construction in Conlisk (2001).

Our next result gives a lower bound to the fitness obtained in NSCs. Let $\underline{M}$ be the pure maxmin value of the underlying game:

$$\underline{M} = \max_{a_1 \in A} \min_{a_2 \in A} \pi(a_1, a_2).$$

The pure maxmin value $\underline{M}$ is the minimal fitness payoff a player can guarantee herself in the sequential game in which she plays first, and the opponent replies in an arbitrary way (i.e. not necessarily maximising the opponent's fitness.)

Proposition 3 shows that the pure maxmin value is a lower bound on the fitness payoff obtained in an NSC. The intuition is that if the payoff is lower, then a mutant of cognitive level 1, with preferences such that the maxmin action $a_{\underline{M}}$ is dominant, will outperform the incumbents.

Proposition 3. *If (μ^*, b^*) is an NSC then $\Pi(\mu^*, b^*) \geq \underline{M}$.*

Proof. Assume to the contrary that $\Pi(\mu^*, b^*) < \underline{M}$. Let $a_{\underline{M}}$ be a maxmin action of a player, which guarantees that the player's payoff is at least $\underline{M}$, i.e. $a_{\underline{M}} \in \arg\max_{a_1 \in A} \min_{a_2 \in A} \pi(a_1, a_2)$.

Let $u^{a_{\underline{M}}}$ be the preferences in which the player obtains a payoff of 1 if she plays $a_{\underline{M}}$ and a payoff of 0 otherwise. Consider a monomorphic group of mutants with type $(u^{a_{\underline{M}}}, 1)$. The fact that $a_{\underline{M}}$ is a maxmin action implies that $\pi_{\left(u^{a_{\underline{M}}},1\right)}\left(\tilde{\mu},\tilde{b}\right) \geq \underline{M}$ in any post-entry configuration. Furthermore, due to continuity it holds that $\Pi_\theta\left(\tilde{\mu},\tilde{b}\right) < \underline{M}$ for any $\theta \in C(\mu)$ in all sufficiently close focal post-entry configurations. This contradicts that μ^* is an NSS in $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, and thus it contradicts that (μ^*, b^*) is an NSC. □

We conclude by demonstrating that the lower bound of the maxmin payoff is binding. Specifically, Example 1 shows an NSC in a zero-sum game in which the fitness of the incumbents is arbitrarily close to the lowest feasible payoff in the underlying game -1 (which is equal to the maxmin payoff).

Example 1. Consider the Rock-Paper-Scissors game described above. Assume that $k_2 = 1$, $k_3 > 2$, and $q(2,1) = 1$. For each $\epsilon \in (0,1)$, consider a population in which ϵ of the agents have cognitive level 1, and the remaining $1-\epsilon$ of the agents have level 2. The agents' behaviour is according to the behaviour described in Proposition 2, i.e.: (1) an agent of level 2 deceives a level-1 opponent into taking a pure action that the level-2 agent then best-replies to; thus the level-2 agent earns 1 and her opponent earns -1; and (2) individuals of the same cognitive level play the unique Nash equilibrium, and obtain a payoff of zero in the underlying game. When one takes into account the cognitive cost $k_2 = 1$ of the level-2 agents, this behaviour implies that all incumbents obtain a fitness of $\epsilon - 1$. An analogous argument to the proof of Proposition 2 implies that this configuration is an NSC.

4 Extensions

4.1 Partial Observability When There Is No Deception

As mentioned above, our basic model assumes perfect observability, and Nash equilibrium behaviour, in matches without deception. In what follows we briefly describe the results of a robustness check that relaxes the first of these two assumptions. For brevity, we detail the full technical analysis in Appendix D.

Specifically, we follow Dekel, Ely, and Yilankaya (2007) and assume that in matches without deception, each player privately observes the opponent's type with an exogenous probability p, and with the remaining probability observes an uninformative signal. This general model extends both our baseline model (where $p = 1$) and Dekel, Ely, and Yilankaya's (2007) model (which can be viewed as assuming arbitrarily high deception costs).

The main results of the baseline model ($p = 1$) show that (1) only efficient profiles can be NSCs, and (2) there exist non-Nash efficient NSCs, provided that the cost of deception is sufficiently

large. Our analysis shows that the former result (namely, stability implies efficiency) is robust to the introduction of partial observability: (1) a somewhat weaker notion of efficiency is satisfied by the behaviour of the incumbents with the highest cognitive level in any NSC for any $p > 0$, and (2) in games such as the Prisoner's Dilemma, we show that only the efficient profile can be the outcome of an NSC.

On the other hand, our analysis shows that our second main result (namely, the stability of non-Nash efficient outcomes) is not robust to the introduction of partial observability. Specifically, we show that: (1) non-Nash efficient profiles cannot be NSC outcomes for any $p < 1$ in games like the Prisoner's Dilemma, even when the effective cost of deception is arbitrarily large; and (2) non-Nash efficient outcomes cannot be pure NSC outcomes in all games. If a game admits a profile that is both efficient and Nash, then the profile is an NSC outcome for any $p \in [0, 1]$. If the underlying game does not admit such a profile, then our results show that the environment does not admit a pure NSC for any $p \in (0, 1)$, and that games like the Prisoner's Dilemma do not admit any NSC. This suggests that in order to study stability in such environments one might need to apply weaker solution concepts or to follow a dynamic (rather than static) approach.

4.2 Interdependent Preferences

In the main text we deal exclusively with preferences that are defined only over action profiles. In what follows we briefly describe how to extend the analysis to interdependent preferences, i.e. preferences that may also depend on the opponent's type. A detailed formal analysis is presented in Appendix B. Herold and Kuzmics (2009) study a similar setup while assuming perfect observability of types among all individuals. Their key result is that any mixed action that gives each player a payoff above her maxmin payoff can be the outcome of a stable configuration.[23]

Our main result for interdependent preferences in our setup shows that a pure configuration is stable essentially iff: (1) all incumbents have the same cognitive level n, (2) the cost of level n is smaller than the difference between the incumbents' (fitness) payoff and the minmax/maxmin value, and (3) the deviation gain is smaller than the effective cost of deception against an opponent with cognitive level n. In particular, if the marginal effective cost of deception is sufficiently small, then only Nash equilibria can be the outcomes of pure stable configurations, while if the effective cost of deceiving some cognitive level n is sufficiently high (while the cost of achieving level n is sufficiently low), then essentially any action profile is the outcome of a pure stable configuration

[23] Herold and Kuzmics (2009) expand the framework of Dekel, Ely, and Yilankaya (2007) to include interdependent preferences, i.e. preferences that depend on the opponent's preference type. Under perfect or almost perfect observability, if all preferences that depend on the opponent's type are considered, then any symmetric outcome above the minmax material payoff is evolutionarily stable. In our setting a pure profile also has to be a Nash equilibrium in order to be the sole outcome supported by evolutionarily stable preferences. Herold and Kuzmics (2009) find that non-discriminating preferences (including selfish materialistic preferences) are typically not evolutionarily stable on their own. By contrast, certain preferences that exhibit discrimination are evolutionarily stable. Similarly, evolutionary stability requires the presence of discriminating preferences also in our setup.

(similar to the result of Herold and Kuzmics, 2009, in the setup without deception).

The last part of Appendix B characterises stable configurations in the Hawk-Dove game. We show that such games admit heterogeneous stable configurations in which players with different levels coexist, each type has preferences that induce cooperation only against itself, and higher types "exploit" lower types (and this is offset by their higher cognitive cost).

5 Conclusion and Directions for Future Research

We have developed a model in which preferences coevolve with the ability to detect others' preferences and misrepresent one's own preferences. To this end, we have allowed for heterogeneity with respect to costly cognitive ability. The assumption of an exogenously given level of observability of the opponent's preferences, which has characterised the indirect evolutionary approach up until now, is replaced by the Machiavellian notion of deception equilibrium, which endogenously determines what each player observes. Our model assumes a very powerful form of deception. This allows us to derive sharp results that clearly demonstrate the effects of endogenising observation and introducing deception. We think that the "Bayesian" deception is an interesting model for future research: each incumbent type is associated with a signal, agents with high cognitive levels can mimic the signals of types with lower cognitive levels, and agents maximise their preferences given the received signals and the correct Bayesian inference about the opponent's type.

In a companion paper (Heller and Mohlin, forthcoming) we study environments in which players are randomly matched, and make inferences about the opponent's type by observing her past behaviour (rather than directly observing her type, as is standard in the "indirect evolutionary approach"). In future research, it would be interesting to combine both approaches and allow the observation of past behaviour to be influenced by deception.

Most papers taking the indirect evolutionary approach study the stability of preferences defined over material outcomes. Moreover, it is common to restrict attention to some parameterised class of such preferences. Since we study preferences defined on the more abstract level of action profiles we do not make predictions about whether some particular kind of preferences over material outcomes, from a particular family of utility functions, will be stable or not. It would be interesting to extend our model to such classes of preferences. Furthermore, with preferences defined over material outcomes it would be possible to study coevolution of preferences and deception not only in isolated games, but also when individuals play many different games using the same preferences. We hope to come back to these questions and we invite others to employ and modify our framework in these directions.

A Formal Proofs of Theorems 1 and 2

A.1 Preliminaries

This subsection contains notation and definitions that will be used in the following proofs.

A generous action is an action such that if played by the opponent, it allows a player to achieve the maximal fitness payoff. Formally:

Definition 11. Action $a_g \in A$ is *generous*, if there exists $a \in A$ such that $\pi(a, a_g) \geq \pi(a', a'')$ for all $a', a'' \in A$.

Fix a generous action $a_g \in A$ of the game G. A second-best generous action is an action such that if played by the opponent, it allows a player to achieve the fitness payoff that is maximal under the constraint that the opponent is not allowed to play the generous action a_g. Formally:

Definition 12. Action $a_{g_2} \in A$ is *second-best generous*, conditional on $a_g \in A$ being first-best generous, if there exists $a \in A$ such that $\pi(a, a_{g_2}) \geq \pi(a', a'')$ for all $a', a'' \in A$ such that $a'' \neq a_g$.

Fix a generous action $a_g \in A$, and fix a second-best generous action $a_{g_2} \in A$, conditional on $a_g \in A$ being first-best generous. For each $\alpha \geq \beta \geq 0$, let $u_{\alpha,\beta}$ be the following utility function:

$$u_{\alpha,\beta}(a, a') = \begin{cases} \alpha & a' = a_g \\ \beta & a' = a_{g_2} \\ 0 & \text{otherwise.} \end{cases}$$

Observe that such a utility function $u_{\alpha,\beta}$ satisfies:

1. *Indifference:* the utility function only depends on the opponent's action; i.e. the player is indifferent between any two of her own actions.

2. *Pro-generosity:* the utility is highest if the opponent plays the generous action, second-highest if the opponent plays the second-best generous action, and lowest otherwise.

Let $U_{GI} = \{u_{\alpha,\beta} | \alpha \geq \beta \geq 0\}$ be the family of all such preferences, called *pro-generous indifferent preferences*. Note that U_{GI} includes a continuum of different utilities (under the assumption that G includes at least three actions). Thus, for any set of incumbent types, we can always find a utility function in U_{GI} that does not belong to any of the current incumbents.

A.2 Proof of Theorem 1 (Behaviour of the Highest Types)

A.2.1 Proof of Theorem 1, Part 1

Assume to the contrary that $\pi\left(b_{\bar{\theta}}^N\left(\bar{\theta}\right), b_{\bar{\theta}}^N\left(\bar{\theta}\right)\right) < \hat{\pi}$. (Note that the definition of $\hat{\pi}$ implies that the opposite inequality is impossible.) Let $a_1, a_2 \in A$ be any two actions such that (a_1, a_2) is an

efficient action profile, i.e. $0.5 \cdot (\pi(a_1, a_2) + \pi(a_1, a_2)) = \hat{\pi}$. Let $\theta_1, \theta_2, \theta_3$ be three types that satisfy the following conditions: (1) the types are not incumbents: $\theta_1, \theta_2, \theta_3 \notin C(\mu^*)$, (2) the types have the highest incumbent cognitive level: $n_{\theta_1} = n_{\theta_2} = n_{\theta_3} = \bar{n}$, and (3) the types have different pro-generosity indifferent preferences; $u_{\theta_1}, u_{\theta_2}, u_{\theta_3} \in U_{GI}$ and $u_{\theta_i} \neq u_{\theta_j}$ for each $i \neq j \in \{1, 2, 3\}$. Let μ' be the distribution that assigns mass $\frac{1}{3}$ to each of these types. The post-entry type distribution is $\tilde{\mu} = (1 - \epsilon) \cdot \mu^* + \epsilon \cdot \mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality: $\tilde{b}^N_\theta(\theta') = b^N_\theta(\theta')$ and $\tilde{b}^D_\theta(\theta') = b^D_\theta(\theta')$ for each incumbent pair $\theta, \theta' \in C(\mu^*)$.

2. The mutants play fitness-maximising deception equilibria against incumbents with lower cognitive levels: $\left(\tilde{b}^D_{\theta_i}(\theta'), \tilde{b}^D_{\theta'}(\theta_i)\right) \in FMDE(\theta_i, \theta')$ for each $i \in \{1, 2, 3\}$ and $\theta' \in C(\mu^*)$ with $n_{\theta'} < \bar{n}$. Note that $FMDE(\theta_i, \theta')$ is nonempty in virtue of the construction of U_{GI}.

3. In matches without deception between mutants and incumbents, the mutants mimic $\bar{\theta}$ and the incumbents play the same way they play against $\bar{\theta}$: $\left(\tilde{b}^N_{\theta_i}(\theta'), \tilde{b}^N_{\theta'}(\theta_i)\right) = \left(b^N_{\bar{\theta}}(\theta'), b^N_{\theta'}\left(\bar{\theta}\right)\right)$, for each $i \in \{1, 2, 3\}$ and $\theta' \in C(\mu^*)$.

4. Two mutants of *different* types play efficiently when meeting each other: $\tilde{b}^N_{\theta_i}\left(\theta_{(i+1) \bmod 3}\right) = a_1$ and $\tilde{b}^N_{\theta_i}\left(\theta_{(i-1) \bmod 3}\right) = a_2$ for each $i \in \{1, 2, 3\}$.

5. When two mutants of the *same* type meet, they play the same way $\bar{\theta}$ plays against itself: $\tilde{b}^N_{\theta_i}(\theta_i) = b^N_{\bar{\theta}}\left(\bar{\theta}\right)$ for each $i \in \{1, 2, 3\}$.

In virtue of point 1 the construction $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration (with respect to (μ^*, b^*)). By points 2 and 3 each mutant θ_i earns weakly more than $\bar{\theta}$ against all incumbent types. By points 4 and 5 each mutant earns strictly more than $\bar{\theta}$ against the mutants. In total the average fitness earned by each mutant is strictly higher than that of $\bar{\theta}$, against a population that follows $\left(\tilde{\mu}, \tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, which implies that μ^* is not an NSC.

A.2.2 Proof of Theorem 1, Part 2

Assume to the contrary that $\left(\left(b^D_{\bar{\theta}}(\underline{\theta}), b^D_{\underline{\theta}}\left(\bar{\theta}\right)\right)\right) \notin FMDE\left(\bar{\theta}, \underline{\theta}\right)$. Let $\hat{\theta}$ be a type that satisfies the conditions of: (1) not being an incumbent: $\hat{\theta} \notin C(\mu^*)$, (2) having the highest incumbent cognitive level: $n_{\hat{\theta}} = \bar{n}$, and (3) having pro-generous indifferent preferences: $u_{\hat{\theta}} \in U_{GI}$. Let μ' be the distribution that assigns mass one to type $\hat{\theta}$. The post-entry type distribution is $\tilde{\mu} = (1 - \epsilon) \cdot \mu^* + \epsilon \cdot \mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality: $\tilde{b}^N_\theta(\theta') = b^N_\theta(\theta')$ and $\tilde{b}^D_\theta(\theta') = b^D_\theta(\theta')$ $\forall \theta, \theta' \in C(\mu^*)$.

2. In matches with deception between mutants and incumbents , behaviour is such that the mutants maximise their fitness: $\left(\tilde{b}_{\hat{\theta}}^{D}\left(\theta'\right),\tilde{b}_{\theta'}^{D}\left(\hat{\theta}\right)\right)\in FMDE\left(\hat{\theta},\theta'\right)$ for each $\theta'\in C\left(\mu^{*}\right)$ with $n_{\theta'}<\bar{n}$.

3. In matches without deception between mutants and incumbents, the mutants mimic $\bar{\theta}$ and the incumbents play the same way they play against $\bar{\theta}$: $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\theta'\right),\tilde{b}_{\theta'}^{N}\left(\hat{\theta}\right)\right)=\left(b_{\bar{\theta}}^{N}\left(\theta'\right),b_{\theta'}^{N}\left(\bar{\theta}\right)\right)$, for each $\theta'\in C\left(\mu^{*}\right)$.

4. The mutant $\hat{\theta}$ plays against itself the same way $\bar{\theta}$ plays against itself: $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\hat{\theta}\right),\tilde{b}_{\hat{\theta}}^{N}\left(\hat{\theta}\right)\right)=\left(\tilde{b}_{\bar{\theta}}^{N}\left(\bar{\theta}\right),\tilde{b}_{\bar{\theta}}^{N}\left(\bar{\theta}\right)\right)$.

Note that $\left(\tilde{\mu},\tilde{b}\right)$ is a focal configuration (with respect to (μ^{*},b^{*})) and that $\hat{\theta}$ obtains a strictly higher fitness than $\bar{\theta}$ against a population that follows $\left(\tilde{\mu},\tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^{*} in the population game $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$. Thus, μ^{*} is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, which implies that μ^{*} is not an NSC.

A.2.3 Proof of Theorem 1, Part 3

Assume to the contrary that $\pi\left(\underline{\theta},\bar{\theta}\right)>\hat{\pi}$, which immediately implies that $\pi\left(\bar{\theta},\underline{\theta}\right)<\hat{\pi}$ and that either $\pi\left(b_{\underline{\theta}}^{|D}\left(\bar{\theta}\right),b_{\bar{\theta}}^{D}\left(\underline{\theta}\right)\right)>\hat{\pi}$ or $\pi\left(b_{\underline{\theta}}^{N}\left(\bar{\theta}\right),b_{\bar{\theta}}^{N}\left(\underline{\theta}\right)\right)>\hat{\pi}$. Let $\hat{\theta}$ be a type that satisfies the conditions of: (1) not being an incumbent: $\hat{\theta}\notin C\left(\mu^{*}\right)$, (2) having the highest incumbent cognitive level: $n_{\hat{\theta}}=\bar{n}$, and (3) having pro-generous indifferent preferences: $u_{\hat{\theta}}\in U_{GI}$. Let μ' be the distribution that assigns mass one to type $\hat{\theta}$. The post-entry type distribution is $\tilde{\mu}=(1-\epsilon)\cdot\mu^{*}+\epsilon\cdot\mu'$. Let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality: $\tilde{b}_{\theta}^{N}\left(\theta'\right)=b_{\theta}^{N}\left(\theta'\right)$ and $\tilde{b}_{\theta}^{D}\left(\theta'\right)=b_{\theta}^{D}\left(\theta'\right)\forall\theta,\theta'\in C\left(\mu^{*}\right)$.

2. In matches with deception between mutants and incumbents, behaviour is such that the mutants maximise their fitness: $\left(\tilde{b}_{\hat{\theta}}^{D}\left(\theta'\right),\tilde{b}_{\theta'}^{D}\left(\hat{\theta}\right)\right)\in FMDE\left(\hat{\theta},\theta'\right)$ for each $\theta'\in C\left(\mu^{*}\right)$ with $n_{\theta'}<\bar{n}$.

3. In a match between a mutant $\hat{\theta}$ and the incumbent $\bar{\theta}$, the mutant mimics $\underline{\theta}$, and the incumbent $\bar{\theta}$ plays the same way it plays against $\underline{\theta}$: $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\bar{\theta}\right),\tilde{b}_{\bar{\theta}}^{N}\left(\hat{\theta}\right)\right)=\left(b_{\underline{\theta}}^{N}\left(\bar{\theta}\right),b_{\bar{\theta}}^{N}\left(\underline{\theta}\right)\right)$ if $\pi\left(b_{\underline{\theta}}^{N}\left(\bar{\theta}\right),b_{\bar{\theta}}^{N}\left(\underline{\theta}\right)\right)>\hat{\pi}$, and $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\bar{\theta}\right),\tilde{b}_{\bar{\theta}}^{N}\left(\hat{\theta}\right)\right)=\left(b_{\underline{\theta}}^{D}\left(\bar{\theta}\right),b_{\bar{\theta}}^{D}\left(\underline{\theta}\right)\right)$ otherwise.

4. The mutant $\hat{\theta}$ plays against itself the same way $\bar{\theta}$ plays against itself: $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\hat{\theta}\right),\tilde{b}_{\hat{\theta}}^{N}\left(\hat{\theta}\right)\right)=\left(\tilde{b}_{\bar{\theta}}^{N}\left(\bar{\theta}\right),\tilde{b}_{\bar{\theta}}^{N}\left(\bar{\theta}\right)\right)$.

5. The mutant $\hat{\theta}$ mimics $\bar{\theta}$ against all other incumbents without deception, and these incumbents play against $\hat{\theta}$ in the same way they play against $\bar{\theta}$: $\left(\tilde{b}_{\hat{\theta}}^{N}\left(\theta'\right),\tilde{b}_{\theta'}^{N}\left(\hat{\theta}\right)\right)=\left(b_{\bar{\theta}}^{N}\left(\theta'\right),b_{\theta'}^{N}\left(\bar{\theta}\right)\right)$ for each $\theta'\neq\bar{\theta}$.

Note that $\left(\tilde{\mu},\tilde{b}\right)$ is a focal configuration (with respect to (μ^*,b^*)). By point 2 the mutant $\hat{\theta}$ earns weakly more than $\bar{\theta}$ against lower types. By point 3 and Theorem 1.1, the mutants earn strictly more than $\bar{\theta}$ against type $\bar{\theta}$. By points 3 and 4 and Theorem 1.1, the mutant earns strictly more than $\bar{\theta}$ against the mutant. By point 5 the mutant $\hat{\theta}$ earns the same as $\bar{\theta}$ against all other types. In total the average fitness earned by $\hat{\theta}$ is strictly higher than that of $\bar{\theta}$, against a population that follows $\left(\tilde{\mu},\tilde{b}\right)$. Recall (Remark 4 in Section 2.3) that all the incumbent types have the same fitness in (μ^*,b^*). By a standard continuity argument, the fitness of incumbent $\bar{\theta}$ is arbitrarily close (for a sufficiently small ϵ) to the fitness levels of any other incumbent type in the focal post-entry configuration $\left(\tilde{\mu},\tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the type game $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS, in $\Gamma_{\left(\tilde{\mu},\tilde{b}\right)}$, which implies that (μ^*,b^*) is not an NSC.

A.3 Proof of Case (A) in Theorem 2

In what follows we fill in the missing technical details for the part of the proof of Theorem 2 that concerns case (A). We begin by proving a lemma.

Lemma 2. *If $(\sigma_1,\sigma_2)\in DE(\theta_1,\theta_2)$ then there exist actions $a_1,a_1'\in C(\sigma_1)$ and $a_2,a_2'\in C(\sigma_2)$ such that: $(a_1,a_2)\in DE(\theta_1,\theta_2)$, and $(a_1',a_2')\in DE(\theta_1,\theta_2)$, with $\pi(a_1,a_2)\geq\pi(\sigma_1,\sigma_2)$, and $\pi(a_1',a_2')\leq\pi(\sigma_1,\sigma_2)$.*

Proof. Note that for any mixed deception equilibrium (σ_1,σ_2) and any action $a\in C(\sigma_2)$, the profile (σ_1,a) is also a deception equilibrium (because otherwise the deceiver would not induce the deceived party to take a mixed action that puts positive weight on a). It follows that there are actions $a_2,a_2'\in C(\sigma_2)$ such that (σ_1,a_2) and (σ_1,a_2') are deception equilibria, with $\pi(\sigma_1,a_2)\geq\pi(\sigma_1,\sigma_2)$ and $\pi(\sigma_1,a_2')\leq\pi(\sigma_1,\sigma_2)$. Furthermore, if (σ_1,a_2) and (σ_1,a_2') are deception equilibria, then for any action $a\in C(\sigma_1)$, the profiles (a,a_2) and (a,a_2') are also deception equilibria, with $\pi(\sigma_1,a_2)=\pi(a,a_2)$ and $\pi(\sigma_1,a_2')=\pi(a,a_2')$. Hence there are actions $a_1,a_1'\in C(\sigma_1)$ such that (a_1,a_2) and (a_1',a_2') are deception equilibria, with $\pi(a_1,a_2)=\pi(\sigma_1,a_2)\geq\pi(\sigma_1,\sigma_2)$, and $\pi(a_1,a_2')=\pi(\sigma_1,a_2')\leq\pi(\sigma_1,\sigma_2)$. □

Assume that case (A) holds: there is an incumbent $\mathring{\theta}$ that plays inefficiently against itself, i.e. $\left(b^N_{\mathring{\theta}}\left(\mathring{\theta}\right),b^N_{\mathring{\theta}}\left(\mathring{\theta}\right)\right)\neq(\bar{a},\bar{a})$, and there is no incumbent type with a strictly higher cognitive level than $\mathring{\theta}$ that satisfies any of the cases (A), (B), or (C). To prove that this cannot hold in an NSC we introduce a mutant $\hat{\theta}=(\hat{u},n_{\mathring{\theta}})\notin C(\mu^*)$. If $\Sigma(u_{\mathring{\theta}})=\Delta$, then we let $\hat{u}\in U_{GI}$ be such that $\hat{\theta}=(\hat{u},n_{\mathring{\theta}})\notin C(\mu^*)$. If $\Sigma(u_{\mathring{\theta}})\neq\Delta$, then we fix a dominated action $\underline{a}\in A\backslash\Sigma(u_{\mathring{\theta}})$, and let $\hat{u}$ be defined as follows:

$$\hat{u}(a,a')=\begin{cases}max_{a\in A}\left(u_{\mathring{\theta}}(a,\bar{a})\right) & a=a'=\bar{a}\\ u_{\mathring{\theta}}(\underline{a},a')-\beta_{a'} & a=\underline{a}\ and\ a'\neq\bar{a}\\ u_{\mathring{\theta}}(a,a') & \text{otherwise,}\end{cases}$$

where each $\beta_{a'} \geq 0$ is chosen such that $\hat{\theta} = (\hat{u}, n_{\mathring{\theta}}) \notin C(\mu^*)$. That is, if $\Sigma(u_{\mathring{\theta}}) \neq \Delta$, then the utility function $\hat{u}$ is constructed from the utility function $u_{\mathring{\theta}}$ by arbitrarily lowering the payoff of some of the outcomes associated with the (already) dominated action $\underline{a}$ and that do not involve action $\bar{a}$, while increasing the payoff of the outcome $(\bar{a}, \bar{a})$ by the minimal amount that makes $\bar{a}$ a best reply to itself. Note that this definition of $\hat{u}$ is valid also for the case of $\bar{a} = \underline{a}$. It follows that $a \in \Sigma(u_{\mathring{\theta}}) \cup \{\bar{a}\}$ iff $a \in \Sigma(\hat{u})$. To see this, note that if $\Sigma(u_{\mathring{\theta}}) \neq \Delta$ and $\underline{a} = \bar{a}$, then $\Sigma(\hat{u}) = \Sigma(u_{\mathring{\theta}}) \cup \{\bar{a}\}$. Otherwise $\Sigma(\hat{u}) = \Sigma(u_{\mathring{\theta}})$. Thus, $\hat{\theta}$ can be induced to play exactly the same pure actions as $\mathring{\theta}$, unless $\bar{a} = \underline{a}$, in which case $\hat{\theta}$ can be induced to play $\bar{a}$ in addition to all actions that $\mathring{\theta}$ can be induced to play.

Let μ' be the distribution that assigns mass one to type $(\hat{u}, n_{\mathring{\theta}})$. Let the post-entry type distribution be $\tilde{\mu} = (1-\epsilon) \cdot \mu^* + \epsilon \cdot \mu'$, and let the post-entry behaviour policy $\tilde{b}$ be defined as follows:

1. Behaviour among incumbents respects focality: $\tilde{b}^N_\theta(\theta') = b^N_\theta(\theta')$ and $\tilde{b}^D_\theta(\theta') = b^D_\theta(\theta')$ $\forall \theta, \theta' \in C(\mu^*)$.

2. In matches without deception between the mutant type $\hat{\theta}$ and any incumbent type θ', the mutant $\hat{\theta}$ mimics $\mathring{\theta}$, and the incumbent θ' treats the mutant $\hat{\theta}$ like the incumbent $\mathring{\theta}$: $\left(\tilde{b}^N_{\hat{\theta}}(\theta'), \tilde{b}^N_{\theta'}\left(\hat{\theta}\right)\right) = \left(b^N_{\mathring{\theta}}(\theta'), b^N_{\theta'}\left(\mathring{\theta}\right)\right)$ for all θ' such that $n_{\theta'} = n_{\mathring{\theta}}$ and $\theta' \neq \hat{\theta}$.

3. In matches with deception between the mutant type $\hat{\theta}$ and any lower type $\theta' \in C(\mu^*)$ (with $n_{\theta'} < n_{\hat{\theta}}$), we distinguish two cases.

   (a) Suppose that $\Sigma(u_{\mathring{\theta}}) = \Delta$. In this case let $\left(\tilde{b}^D_{\hat{\theta}}(\theta'), \tilde{b}^D_{\theta'}\left(\hat{\theta}\right)\right) \in FMDE\left(\hat{\theta}, \theta'\right)$. Note that $FMDE\left(\hat{\theta}, \theta'\right)$ is nonempty since in this case $\hat{u} \in U_{GI}$.

   (b) Suppose that $\Sigma(u_{\mathring{\theta}}) \neq \Delta$. In this case let $\left(\tilde{b}^D_{\hat{\theta}}(\theta'), \tilde{b}^D_{\theta'}\left(\hat{\theta}\right)\right) = (a_1, a_2)$, for some $(a_1, a_2) \in DE\left(\mathring{\theta}, \theta'\right)$ such that $\pi(a_1, a_2) \geq \pi\left(b^D_{\mathring{\theta}}(\theta'), b^D_{\theta'}\left(\mathring{\theta}\right)\right)$. By Lemma 2 above such a profile (a_1, a_2) exists.

4. The mutant plays efficiently when meeting itself: $\tilde{b}^N_{\hat{\theta}}\left(\hat{\theta}\right) = \bar{a}$.

5. In matches with deception between the mutant $\hat{\theta}$ and a higher type $\theta' \in C(\mu^*)$ $(n_{\theta'} > n_{\hat{\theta}})$, we distinguish two cases. Pick a profile $(a_1, a_2) \in DE\left(\theta', \mathring{\theta}\right)$, such that $\pi(a_2, a_1) \geq \pi\left(b^D_{\mathring{\theta}}(\theta'), b^D_{\theta'}\left(\mathring{\theta}\right)\right)$. By Lemma 2 above such a profile (a_1, a_2) exists. Moreover, by the construction of $\hat{u}$, it is either the case that $(a_1, a_2) \in DE\left(\theta', \hat{\theta}\right)$, or there is some $\tilde{a}$ such that $u_{\theta'}(\tilde{a}, \bar{a}) > u_{\theta'}(a_1, a_2)$. In the latter case we have $(\bar{a}, \bar{a}) \in DE\left(\theta', \hat{\theta}\right)$, due to the fact that $\left(b^D_{\theta'}(\theta'), b^D_{\theta'}(\theta')\right) = (\bar{a}, \bar{a})$ implies that $\bar{a}$ is a best reply to $\bar{a}$ for type θ'.

   (a) If $u_{\theta'}(a_1, a_2) > u_{\theta'}(\bar{a}, \bar{a})$ let $\left(\tilde{b}^D_{\theta'}\left(\hat{\theta}\right), \tilde{b}^D_{\hat{\theta}}(\theta')\right) = (a_1, a_2)$. Note that by the definition of (a_1, a_2) it holds that $\pi(a_2, a_1) \geq \pi\left(b^D_{\hat{\theta}}(\theta'), b^D_{\theta'}\left(\mathring{\theta}\right)\right)$.

(b) If $u_{\theta'}(a_1, a_2) \leq u_{\theta'}(\bar{a}, \bar{a})$ let $\left(\tilde{b}^D_{\theta'}\left(\hat{\theta}\right), \tilde{b}^D_{\hat{\theta}}(\theta')\right) = (\bar{a}, \bar{a})$. Note that by the definition of $\mathring{\theta}$ it holds that $\pi(\bar{a}, \bar{a}) \geq \pi\left(b^D_{\mathring{\theta}}(\theta'), b^D_{\theta'}\left(\mathring{\theta}\right)\right)$.

By point 1, $\left(\tilde{\mu}, \tilde{b}\right)$ is a focal configuration (with respect to (μ^*, b^*)). By point 2 the mutant $\hat{\theta}$ earns weakly more than $\mathring{\theta}$ against lower types. By point 3 the mutant $\hat{\theta}$ earns the same as $\mathring{\theta}$ against all incumbents of level $n_{\mathring{\theta}}$. By points 3 and 4 (and the assumption that $\mathring{\theta}$ does not play efficiently against itself), the mutant $\hat{\theta}$ earns strictly more than $\mathring{\theta}$ against $\hat{\theta}$. By point 5 the mutant $\hat{\theta}$ earns weakly more than $\mathring{\theta}$ against all incumbents of a higher cognitive level. In total the average fitness earned by $\hat{\theta}$ is strictly higher than that of $\mathring{\theta}$, against a population that follows $\left(\tilde{\mu}, \tilde{b}\right)$. This implies that μ' is a strictly better reply against μ^* in the population game $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$. Thus, μ^* is not a symmetric Nash equilibrium, and therefore it is not an NSS of $\Gamma_{\left(\tilde{\mu}, \tilde{b}\right)}$, which implies that μ^* is not an NSC. Thus we have shown that $\mathring{\theta}$ plays efficiently against itself.

B Type-interdependent Preferences

As argued by Herold and Kuzmics (2009, pp. 542–543), people playing a game seem to care not only about the outcome, but also their opponent's intentions and they discriminate between different types of opponents (for experimental evidence, see, e.g., Falk, Fehr, and Fischbacher, 2003; Charness and Levine, 2007). Motivated by this observation, in this appendix we extend our baseline model to allow preferences to depend not only on action profiles, but also on an opponent's type.

B.1 Changes to the Baseline Model

We briefly describe how to extend the model to handle type-interdependent preferences. Our construction is similar to that of Herold and Kuzmics (2009).

When the preferences of a type depend on the opponent's type, we can no longer work with the set of all possible preferences, because it would create problems of circularity and cardinality.[24] Instead, we must restrict attention to a pre-specified set of feasible preferences. We begin by defining Θ_{ID} as an arbitrary set of labels. Each label is a pair $\theta = (u, n) \in \Theta_{ID}$, where $n \in \mathbb{N}$ and u is a type-interdependent utility function that depends on the played action profile as well as the opponent's label, $u : A \times A \times \Theta_{ID} \to \mathbb{R}$.

[24] The circularity comes from the fact that each type contains a preferences component, which is identified with a utility function defined over types (and action profiles). To see that this creates a problem if the set of types is unrestricted, let U_* be the set of all utility functions that we want to include in our model. Hence $\Theta_* = U_* \times \mathbb{N}$ is the set of all types. If U_{**} is the set of *all* mappings $u : A \times A \times \Theta_* \to \mathbb{R}$, or, equivalently, U_{**} is the set of *all* mappings $u : A \times A \times U_* \times \mathbb{N} \to \mathbb{R}$, then clearly we have $U_{**} \neq U_*$. See also footnote 10 in Herold and Kuzmics (2009).

Each label $\theta = (u, n)$ may now be interpreted as a type. The definition of u extends to mixed actions in the obvious way. We use the label u also to describe its associated utility function u. Thus $u(\sigma, \sigma', \theta')$ denotes the subjective payoff that a player with preferences u earns when she plays strategy σ against an opponent with type θ' who plays strategy σ'.

Let U_{ID} denote the set of all preferences that are part of some type in Θ_{ID}, i.e. $U_{ID} = \{u : \exists n \in \mathbb{N} \text{ s.t. } (u, n) \in \Theta_{ID}\}$. For each preference $\tilde{u} \in U$ of the baseline model (which is defined only over the action profiles) we can define an equivalent type-interdependent preference $u \in U_{ID}$, which is independent of the opponent's type; that is, $u(\sigma, \sigma', \theta') = u(\sigma, \sigma', \theta'') = \tilde{u}(\sigma, \sigma')$ for each $\theta', \theta'' \in \Theta_{ID}$ and $\sigma, \sigma' \in \Delta(A)$. Let U_N denote the set of all such type-interdependent versions of the preferences of the baseline model. To simplify the statements of the results of Section B.3, in what follows we assume that $U_N \subseteq U_{ID}$.

Next, we amend the definitions of Nash equilibrium, undominated strategies, and deception equilibrium. The best-reply correspondence now takes both strategies and types as arguments: $BR_u(\sigma', \theta') = \arg\max_{\sigma \in \Delta(A)} u(\sigma, \sigma', \theta')$. Accordingly we adjust the definition of the set of Nash equilibria,

$$NE(\theta, \theta') = \{(\sigma, \sigma') \in \Delta(A) \times \Delta(A) : \sigma \in BR_u(\sigma', \theta') \text{ and } \sigma' \in BR_{u'}(\sigma, \theta)\},$$

and the set of *undominated strategies,*

$$\Sigma(\theta) = \{\sigma \in \Delta(A) : \text{there exists } \sigma' \in \Delta(A) \text{ and } \theta' \in \Theta_{ID} \text{ such that } \sigma \in BR_u(\sigma', \theta')\}.$$

Finally, we adapt the definition of deception equilibrium. Given two types θ, θ' with $n_\theta > n_{\theta'}$, a strategy profile $(\tilde{\sigma}, \tilde{\sigma}')$ is a *deception equilibrium* if

$$(\tilde{\sigma}, \tilde{\sigma}') \in \arg \max_{\sigma \in \Delta(A), \sigma' \in \Sigma(\theta')} u_\theta(\sigma, \sigma', \theta').$$

The interpretation of this definition is that the deceiver is able to induce both a belief about the deceiver's preferences, and a belief the deceiver's intention, in the mind of the deceived party. Let $DE(\theta, \theta')$ be the set of all such deception equilibria. The rest of our model remains unchanged.

Some of the following results rely on the existence of preferences $u^{\tilde{a}}_{\tilde{a}', \tilde{n}}$ that satisfy two conditions: (1) action $\tilde{a}$ is a (subjective) dominant action against an opponent with the same preferences and with cognitive level $\tilde{n}$, and (2) action $\tilde{a}'$ is the dominant action against all other opponents. Formally:

Definition 13. Given any two actions $\tilde{a}, \tilde{a}' \in A$, let $u^{\tilde{a}}_{\tilde{a}', \tilde{n}}$ be the discriminating preferences defined

by the following utility function: for all $a, a' \in A$ and $\theta' \in U_{ID}$,

$$u^{\tilde{a}}_{\tilde{a}',\tilde{n}}(a, a', \theta') = \begin{cases} 1 & \left(\theta' = \left(u^{\tilde{a}}_{\tilde{a}',\tilde{n}}, \tilde{n}\right) \text{ and } a = \tilde{a}\right) \text{ or } \left(\theta' \neq \left(u^{\tilde{a}}_{\tilde{a}',\tilde{n}}, \tilde{n}\right) \text{ and } a = \tilde{a}'\right) \\ 0 & \text{otherwise.} \end{cases}$$

Finally, define the *effective cost of deceiving cognitive level* n, denoted by $c(n)$, as the minimal ratio between the additional cognitive cost and the probability of deceiving an opponent of cognitive level n:

$$c(n) = \min_{m>n} \frac{k_m - k_n}{q(m,n)}.$$

Note that $c(1) \equiv c$, which coheres with the definition of the effective cost of deception (with respect to cognitive level 1) in the baseline model.

B.2 Pure Maxmin and Minimal Fitness

The pure maxmin and minmax values give a minimal bound to the fitness of an NSC. Given a game $G = (A, \pi)$, define $\underline{M}$ and $\bar{M}$ as its pure maxmin and minmax values, respectively:

$$\underline{M} = \max_{a_1 \in A} \min_{a_2 \in A} \pi(a_1, a_2), \qquad \overline{M} = \min_{a_2 \in A} \max_{a_1 \in A} \pi(a_1, a_2).$$

The pure maxmin value $\underline{M}$ is the minimal fitness payoff a player can guarantee herself in the sequential game in which she plays first, and the opponent replies in an arbitrary way. The pure minmax value $\overline{M}$ is the minimal fitness payoff a player can guarantee herself in the sequential game in which her opponent plays first an arbitrary action, and she best-replies to the opponent's pure action. It is immediate that $\underline{M} \leq \overline{M}$ and that the minmax value in mixed actions is between these two values.

Let $a_{\underline{M}}$ be a maxmin action of a player; i.e. an action $a_{\underline{M}}$ guarantees that the player's payoff is at least $\underline{M}$, and let $a_{\bar{M}}$ be a minmax action, i.e. an action that guarantees that the opponent's payoff is at most $\bar{M}$:

$$a_{\underline{M}} \in \arg\max_{a_1 \in A} \min_{a_2 \in A} \pi(a_1, a_2), \qquad a_{\bar{M}} \in \arg\min_{a_2 \in A} \max_{a_1 \in A} \pi(a_1, a_2).$$

The proof of Proposition 3 holds with minor changes also in the setup of interdependent preferences (under the assumption that $(u^{a_{\underline{M}}}, 1) \in \Theta_{ID}$), and this implies that the maxmin value is a lower bound on the fitness payoff obtained in an NSC (i.e. if (μ, b) is an NSC then $\Pi(\mu, b) \geq \underline{M}$).

B.3 Characterisation of Pure Stable Configurations

In this subsection we show that, essentially, a pure configuration is stable if and only if (1) all incumbents have the same cognitive level n, (2) the cost of level n is smaller than the difference

between the incumbents' (fitness) payoff and the minmax/maxmin values, and (3) the deviation gain is smaller than the effective cost of deceiving cognitive level n.

We begin by formally stating and proving the necessity claim.

Proposition 4. *If (μ^*, a^*) is a pure NSC then the following holds: (1) if $\theta, \theta' \in C(\mu^*)$ then $n_\theta = n_{\theta'} = n$ for some n, (2) $\pi(a^*, a^*) - \underline{M} \geq k_n$, and (3) $g(a^*) \leq c(n)$.*

Proof.

1. Since all players earn the same game payoff of $\pi(a^*, a^*)$, they must also incur the same cognitive cost, or else the fitness of the different incumbent types would not be balanced (which would contradict the fact that (μ, a^*) is an NSC).

2. Assume to the contrary that $\pi(a^*, a^*) - \underline{M} < k_n$. A mutant of type $(\pi, 1)$ will be able to earn at least $\underline{M}$ against incumbents in any post-entry focal configuration. As the fraction of mutants vanishes the average fitness of mutants is weakly higher than $\underline{M}$, whereas the fitness of the incumbents converges to $\pi(a^*, a^*) - k_n$. Thus, if it were the case that $\pi(a^*, a^*) - \underline{M} < k_n$, then the mutants would outperform the incumbents.

3. Assume to the contrary that $g(a^*) > c(n)$. This implies that there exists a cognitive level $m > n$ such that $g(a^*) > \frac{k_m - k_n}{q(m,n)}$. Let $\tilde{a}$ be the fitness best reply against a^*. Let $\tilde{u} \in U_N$ be the preferences that assign a subjective payoff of one if the agent plays either $\tilde{a}$ or a^* and the opponent plays a^*, and zero otherwise, i.e. $\tilde{u}(a, a', \theta') = \mathbf{1}_{a \in \{a^*, \tilde{a}\} \text{ and } a' = a^*}$. There is a focal post-entry configuration in which all agents play action a^* in all interactions except when a deceiving mutant plays action $\tilde{a}$. A mutant of type $(\tilde{u}, m)$ will then earn $\pi(a^*, a^*) + g(a^*) \cdot q(m, n)$ against the incumbents. As the fraction of mutants vanishes the average fitness of mutants is weakly higher than

$$\pi(a^*, a^*) + g(a^*) \cdot q(m, n) - k_m > \pi(a^*, a^*) + (k_m - k_n) - k_m = \pi(a^*, a^*) - k_n,$$

whereas the fitness of the incumbents is weakly below $\pi(a^*, a^*) - k_n$. Thus, if it were true that $g(a^*) > c(n)$, the mutants would strictly outperform the incumbents.

□

Next, we state and prove the sufficiency claim.

Proposition 5. *Suppose* that $\hat{\theta} := \left(u^{a^*}_{a_{\bar{M},n}}, n\right) \in \Theta_{ID}$. *If $\pi(a^*, a^*) - \overline{M} > k_n$, and $g(a^*) < c(n)$, then $\left(\hat{\theta}, a^*\right)$ is an ESC.*

Proof. Suppose that all incumbents are of type $\left(u^{a^*}_{a_{\bar{M},n}}, n\right)$. Note that in all focal post-entry configurations the incumbent $\hat{\theta}$ always plays either a^* or $a_{\bar{M}}$. Moreover, whenever an incumbent agent is

non-deceived, then she plays action a^* against a fellow incumbent and action $a_{\bar{M}}$ against a mutant. The fact that $\pi(a^*, a^*) - k_n > \overline{M}$ implies that any mutant $\theta \neq \hat{\theta}$ with cognitive level $n_{\theta'} \leq n$ earns a strictly lower payoff against the incumbents in any focal post-entry configuration. As a result, if the frequency of mutants is sufficiently small, then they are strictly outperformed. Against a mutant (θ', n') with cognitive level $n' > n$, an incumbent may play action a^* only when she is being deceived. Since $\pi(a^*, a^*) > \overline{M}$ the mutants earn (on average) at most $\pi(a^*, a^*) + g(a^*) \cdot q(n', n)$ in matches against incumbents. Consequently, as the fraction of mutants vanishes, the average fitness of mutants is weakly less than

$$\pi(a^*, a^*) + g(a^*) \cdot q(n', n) - k_{n'} < \pi(a^*, a^*) + \frac{k_{n'} - k_n}{q(n', n)} \cdot q(n', n) - k_{n'} = \pi(a^*, a^*) - k_n,$$

and the average fitness of the incumbents converges to $\pi(a^*, a^*) - k_n$. Hence, the mutants are outperformed. □

In particular, our results imply that:

1. Any pure equilibrium that induces a payoff above the minmax value $\overline{M}$ is the outcome of a pure ESC (regardless of the cost of deception).

2. If the effective cost of deception is sufficiently small, then only Nash equilibria can be the outcomes of pure NSCs. Specifically, this is the case if $c(n) < g(a)$ for each cognitive level n and each action a such that (a, a) is not a Nash equilibrium of the fitness game.

3. If there is a cognitive level n, such that (1) the cost of achieving level n is sufficiently small, and (2) the effective cost of deceiving an opponent of level n is sufficiently high, then essentially any pure profile is the outcome of a pure ESC (similar to the results of Herold and Kuzmics, 2009, in the setup without deception). Formally, let $A' \subseteq A$ be the set of actions that induce a payoff above the minmax value: $A' = \left\{a \in A | \pi(a, a) > \bar{M}\right\}$. Assume that there is a cognitive level n, such that (1) $k_n < \pi(a, a) - \bar{M}$ for each action $a \in A'$ and (2) $c(n) > g(a)$ for each action a. Then any action $a \in A'$ is the outcome of a pure ESC (in which all incumbents have cognitive level n).

B.4 Application: In-group Cooperation and Out-group Exploitation

The following table represents a family of Hawk-Dove games. When both players play D (Dove) they earn 1 each and when they both play H (Hawk) they earn 0. When a player plays H against an opponent playing D, she obtains an additional gain of $g > 0$ and the opponent incurs a loss of

$l \in (0,1)$.

$$\begin{array}{cc} H & D \\ H \quad 0,0 & 1+g, 1-l \\ D \quad 1-l, 1+g & 1,1 \end{array} . \tag{1}$$

It is natural to think of a mutual play of D as the cooperative outcome. We define preferences that induce players to cooperate with their own kind and to seek to exploit those who are not of their own kind.

Definition 14. Let u^n denote the preferences such that:

1. If $u_{\theta'} = u^n$ and $n_{\theta'} = n$ then $u^n(D, a', \theta') = 1$ and $u^n(H, a', \theta') = 0$ for all a'.
2. If $u_{\theta'} \neq u^n$ or $n_{\theta'} \neq n$ then $u^n(H, a', \theta') = 1$ and $u^n(D, a', \theta') = 0$ for all a'.

Thus, when facing someone who is of the same type, an individual with u^n-preferences strictly prefers cooperation, in the sense of playing D. When facing someone who is not of the same type, an individual with u^n-preferences strictly prefers the aggressive action H.

To simplify the analysis and the notation in this example we assume that a player always succeeds in deceiving an opponent with a lower cognitive level; i.e. we assume that $q(n, n') = 1$ whenever $n > n'$.

Under the assumption that $g > l$ and that the marginal cognitive costs are sufficiently small (but non-vanishing), we construct an ESC in which only individuals with preferences from $\{u^i\}_{i=1}^{\infty}$ are present. Individuals of different cognitive levels coexist, and non-Nash profiles are played in all matches between equals. When individuals of the same level meet, they play mutual cooperation (D, D). When individuals of different levels meet, the higher level plays H and the lower level plays D. The gain from obtaining the high payoff of $1+g$ against lower types is exactly counterbalanced by the higher cognitive costs. By contrast, if $g < l$ then the game does not admit this kind of stable configuration.

Proposition 6. *Let G be the game represented in (1), where $g > 0$ and $l \in (0,1)$. Assume that $q(n, n') = 1$ whenever $n \neq n'$. Suppose that the marginal cognitive cost is small but non-vanishing, so that (a) there is an N such that $k_N \leq l+g < k_{N+1}$, and (b) it holds that $g > k_{n+1} - k_n$ for all $n \leq N$.*

(i) If $g > l$ then there exists an ESC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^n, n)\}_{n=1}^{N}$, and μ^* is mixed (i.e. $|C(\mu^*)| > 1$). The behaviour of the incumbents is as follows: if the individuals in a match are of different cognitive levels, then the higher level plays H and the lower level plays D; if both individuals in a match are of the same cognitive level, then they both play D.

(ii) If $g = l$ then there exists an NSC with the above properties.

(iii) If $g < l$ then there does not exist any NSC (μ^*, b^*), such that $C(\mu^*) \subseteq \{(u^n, n)\}_{n=1}^{\infty}$.

The formal proof is presented in Appendix C.

Remark 8. It is possible to construct an ESC that is like Proposition 6(i) except that when incumbents of the same cognitive level meet they play the mixed equilibrium of the Hawk-Dove game. Thus we can have ESCs in which agents mix at the individual level. For instance, this can be accomplished by considering preferences u^m such that: (1) if $u_{\theta'} = u^m$ and $n_{\theta'} = n$ then $u^m(a, a', \theta') = \pi(a, a', \theta')$ for all a and a', and (2) if $u_{\theta'} \neq u^m$ or $n_{\theta'} \neq n$ then $u^n(H, a', \theta') = 1$ and $u^n(D, a', \theta') = 0$ for all a'.

C Constructions of Heterogeneous NSCs in Examples

Appendix C appears in the supplementary material that can be found online.

D Partial Observability When There Is No Deception

Appendix D appears in the supplementary material that can be found online.

References

Abreu, D., and R. Sethi (2003): "Evolutionary Stability in a Reputational Model of Bargaining," *Games and Economic Behavior*, 44(2), 195–216.

Alger, I., and J. W. Weibull (2013): "Homo Moralis, Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81(6), 2269–2302.

Banerjee, A., and J. W. Weibull (1995): "Evolutionary Selection and Rational Behavior," in *Learning and Rationality in Economics*, ed. by A. Kirman, and M. Salmon. Blackwell, Oxford, pp. 343–363.

Bergstrom, T. C. (1995): "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review*, 85(1), 58–81.

Bester, H., and W. Güth (1998): "Is Altruism Evolutionarily Stable?," *Journal of Economic Behavior and Organization*, 34, 193–209.

Bolle, F. (2000): "Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth," *Journal of Economic Behavior and Organization*, 42, 131–133.

Bomze, I. M., and J. W. Weibull (1995): "Does Neutral Stability Imply Lyapunov Stability?," *Games and Economic Behavior*, 11(2), 173–192.

Brown, G. W., and J. von Neumann (1950): "Solutions of Games by Differential Equations," in *Contributions to the Theory of Games*, ed. by H. W. Kuhn, and A. W. Tucker, Annals of Mathematics Studies 24. Princeton University Press, Princeton.

Byrne, R. W., and A. Whiten (1997): "Machiavellian intelligence," *Machiavellian intelligence II: Extensions and evaluations*, pp. 1–23.

Byrne, R. W., and A. Whiten (1998): *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press, Oxford.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2002): "Sophisticated experience-weighted attraction learning and strategic teaching in repeated games," *Journal of Economic theory*, 104(1), 137–188.

Charness, G., and D. I. Levine (2007): "Intention and stochastic outcomes: An experimental study," *The Economic Journal*, 117(522), 1051–1072.

Conlisk, J. (2001): "Costly Predation and the Distribution of Competence," *American Economic Review*, 91(3), 475–484.

Crawford, V. P. (2003): "Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions," *American Economic Review*, 93(1), 133–149.

Cressman, R. (1997): "Local Stability of Smooth Selection Dynamics for Normal Form Games," *Mathematical Social Sciences*, 34(1), 1–19.

Dekel, E., J. C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economic Studies*, 74, 685–704.

Dufwenberg, M., and W. Güth (1999): "Indirect Evolution vs. Strategic Delegation: A Comparison of Two Approaches to Explaining Economic Institutions," *European Journal of Political Economy*, 15(2), 281–295.

Dunbar, R. I. M. (1998): "The Social Brain Hypothesis," *Evolutionary Anthropology*, 6, 178–190.

Ellingsen, T. (1997): "The Evolution of Bargaining Behavior," *The Quarterly Journal of Economics*, 112(2), 581–602.

Ely, J. C., and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Falk, A., E. Fehr, and U. Fischbacher (2003): "On the nature of fair behavior," *Economic Inquiry*, 41(1), 20–26.

Fershtman, C., and Y. Weiss (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70(1), 53–73.

Frank, R. H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?," *The American Economic Review*, 77(4), 593–604.

Frenkel, S., Y. Heller, and R. Teper (forthcoming): "The endowment effect as a blessing," *International Economic Review*.

Friedman, D., and N. Singh (2009): "Equilibrium Vengeance," *Games and Economic Behavior*, 66(2), 813–829.

Fudenberg, D., and D. K. Levine (1998): *The theory of learning in games*, vol. 2. MIT press.

Gamba, A. (2013): "Learning and Evolution of Altruistic Preferences in the Centipede Game," *Journal of Economic Behavior and Organization*, 85(C), 112–117.

Gauer, F., and C. Kuzmics (2016): "Cognitive empathy in conflict situations," *mimeo*, SSRN 2715160.

Güth, W. (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24(4), 323–344.

Güth, W., and S. Napel (2006): "Inequality Aversion in a Variety of Games: An Indirect Evolutionary Analysis," *The Economic Journal*, 116, 1037–1056.

Güth, W., and M. E. Yaari (1992): "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in *Explaining Process and Change*, ed. by U. Witt. University of Michigan Press, Ann Arbor, MI, pp. 22–34.

Guttman, J. M. (2003): "Repeated Interaction and the Evolution of Preferences for Reciprocity," *The Economic Journal*, 113(489), 631–656.

Heifetz, A., C. Shannon, and Y. Spiegel (2007): "What to Maximize if You Must," *Journal of Economic Theory*, 133(1), 31–57.

Heller, Y. (2015): "Three Steps Ahead," *Theoretical Economics*, 10, 203–241.

Heller, Y., and E. Mohlin (forthcoming): "Observations on cooperation," *Review of Economic studies*.

Herold, F., and C. Kuzmics (2009): "Evolutionary Stability of Discrimination under Observability," *Games and Economic Behavior*, 67, 542–551.

Hines, W. G. S., and J. Maynard Smith (1979): "Games between Relatives," *Journal of Theoretical Biology*, 79(1), 19–30.

Hofbauer, J. (2011): "Deterministic Evolutionary Game Dynamics," in *Proceedings of Symposia in Applied Mathematics*, vol. 69, pp. 61–79.

Hofbauer, J., and K. Sigmund (1988): *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge.

Holloway, R. (1996): "Evolution of the Human Brain," in *Handbook of Human Symbolic Evolution*, ed. by A. Lock, and C. R. Peters. Clarendon Press, New York: Oxford University Press, pp. 74–116.

Hopkins, E. (2014): "Competitive Altruism, Mentalizing and Signalling," *American Economic Journal: Microeconomics*, 6, 272–292.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior*, 28, 13–24.

Humphrey, N. K. (1976): "The Social Function of Intellect," in *Growing Points in Ethology*, ed. by P. P. G. Bateson, and R. A. Hinde. Cambridge University Press, Cambridge, pp. 303-317.

Kim, Y.-G., and J. Sobel (1995): "An Evolutionary Approach to Pre-Play Communication," *Econometrica*, 63(5), 1181–1193.

Kinderman, P., R. I. M. Dunbar, and R. P. Bentall (1998): "Theory-of-Mind Deficits and Causal Attributions," *British Journal of Psychology*, 89, 191–204.

Koçkesen, L., E. A. Ok, and R. Sethi (2000): "Evolution of interdependent preferences in aggregative games," *Games and Economic Behavior*, 31(2), 303–310.

Mailath, G. J., and L. Samuelson (2006): *Repeated games and reputations: long-run relationships*. Oxford university press.

Matsui, A. (1991): "Cheap-Talk and Cooperation in a Society," *Journal of Economic Theory*, 54(2), 245–258.

Maynard Smith, J. (1982): *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J., and G. R. Price (1973): "The Logic of Animal Conflict," *Nature*, 246(5427), 15–18.

Mohlin, E. (2010): "Internalized Social Norms in Conflicts: An Evolutionary Approach," *Economics of Governance*, 11(2), 169–181.

——— (2012): "Evolution of Theories of Mind," *Games and Economic Behavior*, 75(1), 299–312.

Norman, T. W. L. (2012): "Equilibrium Selection and the Dynamic Evolution of Preferences," *Games and Economic Behavior*, 74(1), 311–320.

Ok, E. A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97, 231–254.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior and Organization*, 42, 125–129.

Premack, D., and G. Woodruff (1979): "Does the Chimpanzee Have a Theory of Mind," *Behavioral and Brain Sciences*, 1, 515–526.

Robalino, N., and A. Robson (2016): "The evolution of strategic sophistication," *American Economic Review*, 106(4), 1046–72.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144(3), 379–396.

Robson, A. J. (2003): "The Evolution of Rationality and the Red Queen," *Journal of Economic Theory*, 111, 1–22.

Robson, A. J., and L. Samuelson (2011): "The Evolutionary Foundations of Preferences," in *The Social Economics Handbook*, ed. by J. Benhabib, A. Bisin, and M. Jackson. North Holland, Amsterdam, pp. 221–310.

Rtischev, D. (2016): "Evolution of Mindsight and Psychological Commitment among Strategically Interacting Agents," *Games*, 7(3), 27.

Samuelson, L. (2001): "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, 97(2), 225–230.

Sandholm, W. H. (2001): "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Economic Dynamics*, 4, 637–679.

——— (2010): "Local Stability under Evolutionary Game Dynamics," *Theoretical Economics*, 5(1), 27–50.

Schaffer, M. E. (1988): "Evolutionarily Stable Strategies for a Finite Population and a Variable Contest Size," *Journal of Theoretical Biology*, 132, 469–478.

Schelling, T. C. (1960): *The Strategy of Conflict*. Harvard University Press, Cambrdige, MA.

Schipper, B. C. (2017): "Strategic teaching and learning in games," *mimeo*.

Schlag, K. H. (1993): "Cheap Talk and Evolutionary Dynamics," Bonn Department of Economics Discussion Paper B-242.

Selten, R. (1980): "A Note on Evolutionarily Stable Strategies in Asymmetric Animal Conflicts," *Journal of Theoretical Biology*, 84(1), 93–101.

Sethi, R., and E. Somanthan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

Stahl, D. O. (1993): "Evolution of Smart$_n$ Players," *Games and Economic Behavior*, 5(4), 604–617.

Stennek, J. (2000): "The Survival Value of Assuming Others to be Rational," *International Journal of Game Theory*, 29, 147–163.

Taylor, P. D., and L. B. Jonker (1978): "Evolutionary Stable Strategies and Game dynamics," *Mathematical Biosciences*, 40(1–2), 145–156.

Thomas, B. (1985): "On Evolutionarily Stable Sets," *Journal of Mathematical Biology*, 22(1), 105–115.

van Damme, E. (1987): *Stability and Perfection of Nash Equilibria*. Springer, Berlin.

Wärneryd, K. (1991): "Evolutionary Stability in Unanimity Games with Cheap Talk," *Economics Letters*, 36(4), 375–378.

——— (1998): "Communication, Complexity, and Evolutionary Stability," *International Journal of Game Theory*, 27(4), 599–609.

Weibull, J. W. (1995): *Evolutionary Game Theory*. MIT Press, Cambridge Massachusetts.

Whiten, A., and R. W. Byrne (1988): "Tactical deception in primates," *Behavioral and brain sciences*, 11(2), 233–244.

Wiseman, T., and O. Yilankaya (2001): "Cooperation, secret handshakes, and imitation in the prisoners' dilemma," *Games and Economic Behavior*, 37(1), 216–242.